DOCUMENT RESUME

ED 207 991 TM 810 644

AUTHOR Stone, James C., Ed.; James, Raymond A., Ed.
TITLE Qualitative Evaluation.
INSTITUTION California Council on the Education of Teachers, Berkeley, Calif.
PUB DATE 81
NOTE 141p.
JOURNAL CIT California Journal of Teacher Education; v8 n1 Win 1981

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Evaluation Methods; History; Literary Criticism; Perception; *Program Evaluation; Scientific Methodology
IDENTIFIERS Objectivity; Paradigm Shifts; *Qualitative Research; Quantitative Research; Twin City Institute for Talented Youth MN

ABSTRACT

"Qualitative evaluation" is the theme of this issue of the California Journal of Teacher Education. Ralph Tyler states that evaluation is essentially descriptive, and using numbers does not solve basic problems. Martha Elin Vernazza examines the issue of objectivity in history and its implications for evaluation. She posits that the decisive factor in achieving "objectivity" is the investigator's professional competence in the use of critical judgment. John W. Ratcliffe discusses some recent paradigm shifts in scientific thought, and presents some methodological guidelines for evaluation that are congruent with these shifts. Mark St. John distinguishes between seeing what is happening and one's image of what is happening. Rebecca Zwick and Leonard A. Marascuilo review the evaluation of the Twin City Institute for Talented Youth, and show how it would have been improved with their own quantitative approach. Elliot W. Eisner describes the methodology of educational connoiseurship and educational criticism and illustrates its application. Bruce Joyce and Robert N. Bush describe the California Staff Development Study; and James C. Stone describes an evaluation of Mele-Maat, a British primary school in the New Hebrides. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED207991

CJTE

WINTER
VOLUME 8
NO. 1 1981

CALIFORNIA JOURNAL OF TEACHER EDUCATION

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Stone

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

"Qualitative Evaluation"

TM 810 644

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

CALIFORNIA COUNCIL ON THE EDUCATION OF TEACHERS

CALIFORNIA JOURNAL OF TEACHER EDUCATION

Vol. VIII, No. 1 Winter 1981

Published four times a year by
the California Council on the
Education of Teachers

Theme: QUALITATIVE EVALUATION

Publisher's Note:

The opinions presented in the Journal are those of the authors of the articles and do not necessarily represent those of the Board of Directors of the California Council on the Education of Teachers, the Editorial Board, or the Editor.

RALPH W. TYLER
to whom this issue is dedicated.

General Procedure

The issues of the Journal contain both thematic articles requested from authors by the Guest Editor, who has been selected by the Editorial Board to coordinate that issue, and articles selected by the Panel of Readers. In both instances, the final decision on the articles to be published is the sole responsibility of the Editor. Authors submitting articles for the refereeing process should send them to James C. Stone, Editor, 4607 Tolman Hall, University of California, Berkeley, California 94720. He will forward the article (anonymously) to at least two members of the Panel of Readers. Panelists return the article with their evaluation to the Editor, who forwards a summary to the author. The names of the Readers of any particular manuscript are confidential.

Panel of Readers

David A. Bickimer - Board of Education, City of New York
Gerald J. Brunetti - University of Minnesota
Arthur L. Costa - California State University, Sacramento
Gustavo Gonzales - University of California, Santa Barbara
Sister June Kearney - Holy Names College
Geraldine Larocque - University of North Iowa
Larry L. Leslie - University of Arizona
Jules Pagano - The Massachusetts Regional Community College System, Boston
Larry L. Palmatier - The University of San Francisco
Carol T. Rice - Far West Laboratory for Research and Development, San Francisco
Robert A. Spencer - Webster College
Robert Terrell - University of Missouri, Columbia

Publisher

California Council on the
Education of Teachers

Alvin Thompson, President
James Cusick, Vice President
Claire Pelton, Immediate Past President
James R. Hoffner, Executive Secretary

Directors:
Carol Barnes (1982)
Paul Burke (1983)
Vincent J. Crockenberg (1983)
Philip Fitch (1982)
Theresa Harris (1981)
Alan H. Jones (1983)
Richard D. Jones (1982)
James Neal (1981)
Ruth Schneider (1981)

Editor

James C. Stone
University of California, Berkeley

Guest Editors of this issue

James C. Stone
Raymond A. James

Editorial Board

Arlene Black (1981)
Malcolm P. Douglass (1983)
Walter Hambleton (1983)
James Hoffner, Treasurer
Edna Mitchell (1984)
Ray Roberts (1982)
James C. Stone, Chairman (1983)

Annual subscription rate is $20.00.

Subscriptions and advertisements should be sent to:

Malcolm P. Douglass, Circulation/Production Manager
Harper Hall 203
Claremont Graduate School
Claremont, California 91711

Manuscripts and inquiries should be sent to:

James C. Stone, Editor
4607 Tolman Hall
University of California, Berkeley
Berkeley, California 94720

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording or any information storage and retrieval system, without permission, in writing, from the Editor.

TABLE OF CONTENTS

HOW WE GOT WHERE WE ARE
Ralph W. Tyler 1

OBJECTIVITY IN HISTORY AND EVALUATION
Martha Elin Vernazza 15

NEW PARADIGMS FOR OLD
John W. Ratcliffe 27

TO SEE WHAT IS HAPPENING
Mark St. John 44

CAN EVALUATION PROVIDE USEFUL INFORMATION?
Rebecca Zwick and Leonard A. Marascuilo 61

THE "METHODOLOGY" OF QUALITATIVE EVALUATION: THE CASE OF EDUCATIONAL CONNOISSEURSHIP AND EDUCATIONAL CRITICISM
Elliot W. Eisner 82

THE CALIFORNIA STAFF DEVELOPMENT STUDY
Bruce Joyce and Robert N. Bush 97

MELE-MAAT
James C. Stone 104

Future Issues

Spring, 1981	Refereed Articles Issue
Summer, 1981	BILINGUALISM
Fall, 1981	FACULTY DEVELOPMENT
Winter, 1982	EDUCATION AND THE POLITICAL PROCESS

HOW WE GOT WHERE WE ARE

Ralph W. Tyler

Precis: When this issue of the Journal first was discussed two years ago, the Guest Editors sought out Ralph Tyler, Director Emeritus of the Center for Advanced Study in the Behavioral Sciences, to prepare the lead article. The *Journal* is proud to present an article by the MOST distinguished American educator and scholar of our time, perhaps of all time. Now enjoying his 80th year, Tyler is known rightly as the father of evaluation and the man who first used the word to describe new procedures that he discovered in the late 1930's while working at Ohio State University. At that time testing was the only way to measure school outcomes.

Since the dawn of time, the man has communicated through two sets of symbols--numbers and words. The psychometricians and politicians favor the first. At long last, educators are beginning to recognize that evaluation essentially is a descriptive activity and that, as Tyler says, "Using numbers does not solve basic problems."

It would be a mistake to assume that only in this generation has the problem of conflicting emphasis in qualitative and quantitative evaluation been recognized. Five thousand years ago, the staff of the Chinese emperor debated the value of written examinations as the means of selecting administrators for the new territories being added to the empire as the Chinese armies swept across Asia. The qualifications included competent horsemanship, courage in facing opposition, wisdom in reaching decisions, firmness in administering the edicts, and devotion to the emperor.

Apparently, the process of appraising each candidate on all of the five qualifications was soon discarded. Within a few years the selection procedure followed was to give all candidates the same written examination. Those answers judged to be correct were counted and those candidates who gave the greatest number of correct answers were selected. Why was a qualitative evaluation procedure given up and a simplistic quantitative one adopted? We can only speculate about the causes for the change but it is likely that several factors were influential then as they are now.

Evaluation is Inherently Qualitative

Whether the term evaluation is defined as a process for ascertaining the values in a person, program, or institution or treated as a procedure for comparing the conception or plan with the reality to which it refers, the idea of evaluation is one of qualitative description. In practice, however, the tasks involved are found to be complex and some are difficult. Quantitative treatment at first glance seems to simplify the problems and give the appearance of objectivity and precision that are lacking in qualitative evaluation of persons and institutions.

Evaluation Tasks

To understand more fully the apparent attraction of quantitative evaluation, a review of the tasks involved in a systematic procedure of evaluation is enlightening. The first task is to identify, justify, and define the criteria appropriate for appraising the relevant human functioning or the functioning of the social institution.

A second task is the selection for each criterion of evidence upon which judgments can validly be made. This task involves problems that are widely recognized in the field of psychological and educational testing. It involves (1) validity--what evidence is relevant to the criterion? (2) Reliability--is the evidence available and acceptable to a majority of competent judges or is it a matter of purely idiosyncratic appraisal? (3) Reliability--is the sample of evidence a representative sample of the individual's or institution's relevant functioning, and is it sufficiently adequate in amount to reduce the variations among different samples to a level acceptable for the purposes to be served by the evaluation? (4) Precision--does the treatment of the evidence produce a result that is sufficiently precise for the purposes to be served by the evaluation?

A third task is to produce a synthesis of the judgments made on the several criteria for the individual person or social institution. This synthesis is not necessary when the purpose of the appraisal is to identify strengths and weaknesses of the person or institution, but it has involved serious problems in the past when evaluations were used to select individuals or reject them, or to accredit institutions or to deny accreditation. Decisions of this sort have commonly been made by some sort of summation and and then arranging the individuals or institutions

on an assumed linear continuum. After this arrangement, cutting points are established for acceptance or rejection of the person or institution.

Although they are not always recognized, all evaluations, whether qualitative or quantitative, require the first two tasks and appraisals seeking to sort individuals or institutions rather than merely to describe them require the third task. Throughout the history of the evaluation of teachers and teacher education institutions, the procedures employed have usually been developed without clear recognition of the tasks required and the problems involved. When overlooked, tacit assumptions are made that are often in conflict with the basic purposes and philosophy of the endeavor.

The Influence of Complexity of Teaching

Qualitative evaluation did not appear to be very difficult when teaching was viewed as relatively simple. Even as late as 1910, most leaders of the teaching profession believed that the essential criteria for evaluating the teacher was three: knowing the subject matter to be taught, following an acceptable method of teaching and maintaining classroom discipline. The criteria for accrediting teacher training institutions were also relatively simple, including such standards as: amount of education of the faculty, provision of a "model school", the number of books in the library, and the average annual expenditure per student.

The evaluation of teachers from the Colonial Period until the early part of this century was carried on as a qualitative appraisal. Each teacher was described by the school committee, the principal, teacher or other evaluator in terms of: knowing the subject matter, following an accepted method of teaching, and maintaining class-

room discipline. Teacher training institutions were also described in terms of the accepted criteria rather than being rated or scored in the accreditation process. Although some evaluators used adverbs like very skillful, or extremely knowledgeable, these comments were not given numerical values because during this period there was a continuing shortage of qualified teachers and teacher training institutions. Hence, the purpose of the evaluation process was to identify qualified teachers and institutions. It was not expected to produce scores or ratings that would sharply discriminate among applicants for teaching positions or among institutions seeking accreditation.

This overly simple view of the teaching function was beginning to be questioned after the experiments of Thorndike and others refuted the Doctrine of Formal Discipline, but the publication of the Commonwealth Teacher Training Study in 1928 [1] almost immediately resulted in the recognition by most professional educators that teaching was complex. This investigation of teaching reported 1,001 activities of the American teacher, for too many to be helpfully classified under three categories of functioning. Furthermore, by this time the supply of candidates for teaching positions in the cities exceeded the demand. This meant that the process of teacher evaluation was expected to differentiate among candidates and to indicate the superior ones.

The recognition of the complexity of teaching stimulated the effort to identify a large number of criteria to be used in appraising teachers and institutions, but the agreement among educational leaders as to the important criteria was far from perfect. The lists produced by different writers were quite diverse.

When faced with lack of agreement on qualitative

standards, teacher evaluators sought solutions that avoided comprehensive statements of criteria. At this time, they were greatly impressed by the effectiveness with which the Army Alpha Test had selected candidates for officer training and other specialized military assignments from among the two million draftees in World War I. If a paper-and-pencil test, like the Army Alpha, could evaluate military qualifications, teacher educators argued, such a test could be developed to evaluate teachers. It was not recognized at that time that the Army Alpha Test identified young men who had developed an information background and the learning skills characteristic of educated American youth and did not appraise either innate abilities or special knowledge and skills.

The Cooperative Test Service, in the early 1930s, undertook the development of paper-and-pencil tests designed to sample the content emphasized in the courses offered in programs of teacher education. Although John Flanagan, Director of the Cooperative Test Service, stated in the manual that the National Teacher Examinations appraised only a part of the learning expected of prospective teachers, many educators, untrained in evaluation, assumed that the tests provided a comprehensive measure of teachers and prospective teachers.

These examinations were published at a time when the dual school system of the South was under attack for failing to pay black teachers the same salaries as white teachers when both groups had obtained the bachelor's degree. The average scores on the National Teacher Examinations made by black teachers were found to be much lower than the average scores made by white teachers. This probably accounted for the fact that these examinations were most generally used in southern states. They gave the appearance of objectivity and fairness and were thus used to justify

differentiation in pay. Little attention was given, however, to the fact that the selection of content, the wording of questions, and the weighting of answers are subjective processes.

Because, to the unsophisticated, numbers seem more objective and more precise than descriptive words, qualitative evaluations are given up and quantitative evaluations flourish where accusations of bias and prejudice are frequent.

Another attempt to solve the evaluation problem created by diverse criteria was to assume that professors, supervisors and principals could recognize good teachers and could rank them or rate them in terms of an overall or holistic judgment. If these ratings for a large group of teachers or prospective teachers were obtained, it was posited that the particular criteria of good teaching could be identified by correlating the judgments on the particular criteria with the overall ratings. This procedure requires quantitative overall ratings and quantitative appraisals on particular criteria. The results of this procedure seem valid, objective and precise to those untrained in evaluation. However, the overall ratings are not clear cut appraisals of good teaching, but are based on different, undefined criteria applied to varying samples of evidence by individual judges whose assumptions, biases and prejudices are not known. In addition, the judgments on the particular criteria when transformed into quantitative ratings are distorted by an unmeasured amount in selecting the units and weights used.

An inverse procedure that has been emplyed in seeking to evaluate teachers and institutions is to assemble ratings on many diverse criteria, intercorrelate them, and extract factors and factor loadings. This procedure has the limitations of the overall rating process and in addition is

another step removed from concrete reality by the assumption that the factor solution to a complex matrix has some existence in observable or understandable phenomena.

We are now beginning to recognize that a necessary step in evaluations, whether qualitative or quantitative is the identification, justification and definition of criteria.

Selection of Evidence

When a particular evaluative criterion has been identified, justified and defined, the next step is to select the kinds of evidence to be obtained which are relevant to the criterion. For teaching, much of the evidence is likely to be obtained from teachers, students and the transactions that go on among teachers and students. Observing teachers and students in the classroom would appear to be one of the most valid ways to obtain relevant evidence, but it is often difficult to get a representative sample of classroom transactions and it is time consuming as well. Furthermore, the recording of observations is not easy. Video taping has eased the recording problem but not the logistics nor the sampling problems. It is understandable that written tests, questionnaires and interviews with teachers and students have become the most common means of obtaining data even when their validity as evidence of the criterion is questionable. We are now at a stage in qualitative evaluation in which the definitions of defensible criteria can be used to select kinds of evidence that can clearly furnish a basis for valid judgments.

Objectivity

In the period from 1900 to 1935, American psychologists focused major attention on the problem of

the objectivity of evaluations. Studies of essay examinations found wide variations in the grades assigned by different teachers to the same essays. The short answer form of test which included brief free responses to specific questions as well as true-false and multiple choice exercises could be scored by a key which permitted little if any variation in scores assigned by different judges. This form of test was called an objective test, overlooking the fact that the exercises chosen to represent the criterion, the wording of the exercises, the development of the key and the method of computing total scores involve personal judgments which may be sources of great subjectivity.

The form of English and European examinations were little influenced by the American development of short-answer forms until studies of examinations in the cooperating countries were stimulated and sponsored by the International Conference on Examinations. Then, it was something of a shock to our English and European colleagues to discover wide variations in grades assigned by different examiners in their countries to the same examination papers. It was particularly startling to find that the well-established oral examination procedures in England produced cases in which one examining committee failed a candidate who had been rated "superior" by another examining committee. About the same time, Stalnaker for the College Entrance Examination Board, and Eely for the University of Chicago Board of Examinations both demonstrated that essay examinations could be evaluated with a high degree of objectivity when the evaluative criteria were identified and defined and the judges were given brief training in the evaluation process. Such interesting studies, however, were largely overlooked partly because qualitative evaluations required agreement on standards and were more costly in time and effort.

In the early 1940's, Daniel Prescott and the staff of the National Commission on Teacher Education worked with teachers and prospective teachers using anecdotal records as evidences of their understanding of children. They investigated the objectivity of the ratings of these records and found that this evidence could be rated with a high degree of objectivity when the criteria for the ratings were defined and the raters given several hours of training in their use. Nevertheless, the misconception still exists that quantitative methods are objective while qualitative evaluations involve a great degree of subjectivity.

Reliability

Evidence used as a basis for evaluation is only a sample of the relevant functioning of the person or the institution. Hence, the problem always arises as to the representativeness and the adequacy of the sample, the problem of reliability. A sample of functioning should include the variations found in the day by day functioning of the person or the institution. A representative sample is likely to be obtained if it is randomly chosen or if it is consciously selected to include the known factors likely to make for variations (stratified sampling). An adequate sample is one which includes sufficient cases that its characteristics are quite similar to other samples of equal size. How similar these samples need to be depends on how accurate an estimate of the essential characteristics are required for the purposes the evaluation is to serve.

It is quite possible to estimate the reliability of qualitative evaluations but the methods are not generally known to evaluators of teachers and teacher education instructors. Since 1900, most psychometricians have been preoccupied by the task of sorting, arranging individuals on an assumed continuum. For this purpose, the correlation of

he numerical ratings obtained from two different samples of the same size has furnished an estimate of reliability. This requires quantitative appraisals. The concern with reliability of evaluation has pressed the development of quantitative evaluation while little attention has been given to ascertaining the reliability of particular qualitative evaluations.

Precision

An evaluation can be no more precise than the observations and judgments underlying it. Unfortunately, quantitative evaluations often appear to the unsophisticated as much more precise than they are. For example, recently a great deal was made in the press about the decline of 30 points in the mean scores on Vocabulary made by students taking the Scholastic Aptitude Test. The actual mean decline of so-called raw scores was 2.4 points on a test of some 50 items. But the scoring system of that test translated the raw-score standard deviation into a derived score of 100. The standard deviation was approximately 8 points which by translation became 100 points, and the 2.4 points being three-tenths of the standard deviation became 30 points. Most people did not understand this translation.

Another source of confusion about the precision of a quantitative evaluation is the frequent practice of psychometricians to force the judges to distinguish between two observations that they felt they could not distinguish. The argument for this practice is that when judges are forced to make a distinction, they will furnish more precise judgements. This view has little empirical evidence to support it, but the result of its use is to give apparently precise estimates to what is less precisely appraised. Qualitative evaluations can be stated in as precise terms as the judges are able to make, but like quantitative evaluations, the

results can be no more precise than the judgments on which they are based. If qualitative evaluations are not as precise as desired, efforts should be made to identify the factors involved and to improve the qualitative evaluations rather than to avoid the problem by using quantitative appraisals that only appear to be more precise.

Synthesis

The kind of problem which has appeared to be more difficult to evaluators over the years is that of producing a synthesis of the judgments made on the several criteria for the individual person or institution. To return to the ancient Chinese examination system, it seems to have been adopted partly because the judges did not see how to synthesize the evaluation of a candidate who was judged to be a competent horseman, courageous in facing opposition, lacking wisdom in reaching decisions, wavering in his administration of edicts, and devoted to the emperor. Today, we might well criticize the evidence on which these five evaluations were made, but if we were to accept their validity, objectivity, reliability and precision, we would still have great difficulty in comparing this candidate with one who was a poor horseman, not courageous in facing opposition, wise in reaching decisions, firm in administering edicts and devoted to the emperor. Should the first candidate have preference over the second in selecting administrators?

This problem arises almost wholly in situations where sorting is required such as selecting one or more persons from a larger group. This demands a synthesis to a single scale of some sort. Evaluators have commonly retreated from the very difficult task of making judgment in each case about the relative desirability of different combinations of criteria and have chosen mathematical procedures which appear to solve the

problem with precision. They do not seem to recognize that the selection of a particular mathematical procedure requires the acceptance of certain quantitative relationships within and among criteria.

From time to time, evaluators have sought to solve the problem of syntheses by creating typologies and classifying persons or institutions by placing them in categories of types. For example, while Hitler was ruling Germany, the authoritarian personality was conceived as a type of person in contrast to the democratic and the laissez-faire types. More recently, teachers have been typed as permissive, demanding, and caring. It is possible to describe "ideal types" and work out evaluation procedures for classifying teachers in terms of these types. Two difficulties, however, are commonly encountered. One is that few people can be identified who fit closely the "ideal types." Most people vary somewhat from the type descriptions. In such cases, a quantitative method is developed to define how deviant a person is from the type. Such a method involves assumptions about weights to be given different features of the "ideal type". It also, in practice, assumes that multiple variations from multiple dimensions can be treated as simple linear or geometric functions.

The other difficulty is in explaining the functional value achieved by placing teachers in types. Much effort is usually spent in justifying both the importance and the stability of the typology, and little attention is given to the way in which the use of typologies helps to understand or to improve teaching and learning. Fortunately, we can focus attention on understanding teachers and learning without having to synthesize into one description all the significant evaluations.

In Summary

The past 80 years has given major attention to quantitative evaluation. We are now recognizing that evaluation is essentially descriptive, using numbers does not solve basic problems. This is a time to re-examine qualitative evaluations and to make them better.

REFERENCE CITED

[1]Charters, W. W. and Waples, Douglas, *The Commonwealth Teacher Training Study*, Chicago, University of Chicago Press, 1928.

OBJECTIVITY IN HISTORY AND EVALUATION

Martha Elin Vernazza

Precis: The writer has been an elementary, secondary and college teacher, curriculum consultant and author. Currently she is completing an oral history of Ralph W. Tyler's contribution to program evaluation. In this article she discusses the problem of objectivity in the field of history in comparison with qualitative evaluation. She holds that in historical and evaluation studies, events/happenings/experiences are the "objects" of investigation. Such happenings, however, do not speak for themselves. Vernazza says, "It is the historian or the evaluator who must puzzle out the meanings."

"There is no scientific method as such...The most vital feature of the scientist's procedure has been merely to do the utmost with his mind, no holds barred." [1]

To begin at the beginning as the historian must, the issue of objectivity in history and evaluation is not a question whether history or evaluation is a "science." Rather it is: Upon what grounds are some judgments about a course of events more en-

titled to credence than certain other ones? For this writer, the "grounds" are competence in the professional use of "critical Judgment;" objectivity in history and evaluation is a professional achievement.

As used in this paper, "objectivity" in history and evaluation, or the sense in which phenomena are understood, derives from two professional processes: (1) the observation of external realities in almost the physical sense and (2) participation with the event, situation, person or object of observation through understanding both the cultural and motivational meaning systems "intended" by the participants in a situation. Thus, the sense in which phenomena are understood is a technical or "objective" one. In the process of acquiring it, the professional historian or evaluator is, in effect, developing and using what this writer calls his professional "critical judgment," and in both historical writing and evaluation, it is *competence* in the use of critical judgment that is decisive.

The primary function of the professionals referred to in this paper is to understand the unique event or experience within its broader context of events; for scholarly historical or evaluative understanding, the particular must always relate to something more general. All knowledge that an historian or evaluator has of cultural reality is always knowledge from particular points of view--that is, understandings of how to relate the events of the real world to universal "cultural values" and of how to select those relationships which are significant for us. Through his professional use of critical judgment, the historian or evaluator can attain the perspective and "objectivity" needed to select those elements essential to his professional purposes and leave behind those of his own culture which are irrelevant.

Thus, in aiming for "objectivity," it is essential that both the historian and the evaluator demonstrate their genius through use of critical judgment and interpretation of the event, situation, person, or object of observation. This process of analytical rearrangement of the immediately given reality is key to our comprehension of Reality; the quality of "objectivity" to be achieved depends upon this perpetual reconstruction of concepts. The professional's use of critical judgment builds the essential bridge between cultural levels and those of the concrete actions of individuals.

The First Step: Selection

The first step toward making a judgment in history or in evaluation, therefore, is a process of abstraction which proceeds through the analysis and mental isolation of the components of the given data. In both history and evaluation, this principle of selection might seem too obvious to be considered important. It *is* important, however; its acknowledgment forces our attention to the fact that everything in the writing of history and the process of evaluation depends upon the principle used to control selection. For it is the principle that decides the weight to be assigned to events and ideas, what shall be admitted and what omitted. As well, it decides *how* the selected "facts" are to be arranged and ordered.

We are cultural beings with the capacity and the will to take a deliberate attitude towards the world and to lend it significance (i.e. cultural significance), which will lead us to judge certain phenomena of human existence in its light and to respond to them as being (positively or negatively) meaningful. The analysis of reality is concerned with the configuration into which the selected "facts" are arranged to form a cultural phenomenon significant to us.

The evaluative ideas, which give significance to the fragment of reality analyzed by cultural concepts, are not evaluative in the sense of attributing cultural significance only to "valuable" phenomena. On the contrary. Evaluative ideas identify cultural phenomena *only* because and only insofar as their existence and the form which they historically assume touch in some way on our cultural interests (adapted from Weber, 1949, p.81). Thus, in both history and evaluation, effective critical judgment depends upon the professional's possessing the necessary understanding of the particular cultural and motivational meaning systems involved in a given observation. The professional must determine what the group or "object" of observation he is working with attaches (or, in the case of history, attached) cultural significance to.

Equally important in determining the quality of professional critical judgment is the investigator's differentiation of his role as historian or evaluator from that of simply a participant in his general culture. We see, therefore, that by using the process of critical judgment, the professional investigator is provided determiners for attempting to control his data, and a careful search is conducted for all possibly significant information.

Studying Events, Not Organisms

History and evaluation reports may *seem* most objective, most definite, most uncontroversial in those monographs in which an extremely narrow segment of reality has been explored and reported in a detailed or statistical context. Yet, it is neither the historian's nor the evaluator's primary, or ultimate, function to study events as one might study an organism. To say of history or evaluation that it should approximate to the condition of a science is to ask each to contradict its very

essence. The notion that an historical or professional evaluation inquiry can produce, reinstate, or interpret the events being studied--independent of one another or of an even broader context of events--is incredibly naive. The unique event is comprehensible only in its general context, and for scholarly historical or evaluative understanding the particular must always relate to something more general. Only by analyzing and comparing a wide range of data can the historian or evaluator hope to achieve an understanding of them, for understanding must involve a selection of the evidence and a synthesis on different levels (ranging from literal meaning, observation, reporting of detail, bias, and subjectivity to inference, relevance, and arrangement of ideas) into a meaningful and purposeful account.

Significance of Interpretation

As mentioned above, the historian and the evaluator do not merely *analyze*; they must *synthesize*, and in both fields synthesis occurs, to a certain extent, in the act of analysis itself. This is so because "historical (and evaluation) knowledge is primarily a view of something--much as people passing through a landscape absorb its beauty as they go (citing Huizinga in Stern, 1956).

In this light, we can say that history and evaluation are each a dialogue. For history each age must reinterpret its own past; for evaluation, each situation must reflect on where it is in light of where it has come from and where it wants to go. History and evaluation are each a constant reexamination and reinterpretation; they are also cumulative developments. No ultimately proven synthesis, however, can be reached in either professional process through the sheer addition of "answers" to small questions. The historian's tradition is not so much to look for the perfection of microscopic units of research as it is to try to cope with

certain insistent macroscopic questions (citing Hofstadter in Stern, 1956). The evaluator's tradition is not so much to describe the components of the situation as it is to identify the significant questions whose answers will promote an assessment of merit. As Clark expresses it, "History to mean anything must be more than a rehearsal of facts, it must include an interpretation of the facts." And so it is for an evaluation: to mean anything it must be presented, not as individual statistical measurements of isolated descriptions, but as an integrated whole with positive and meaningful direction. The idea of the continuity of historical and of evaluation knowledge entails this conclusion necessarily. Changes acquire temporal significance only when they are interpreted in terms of a direction *from* something *to* something. The "to" and "from" in question may be intelligently located at any chosen date and place and the limitation is relative to the purpose and problem of the inquiry; it is not inherent in the course of ongoing events (citing Dewey in Meyerhoff, 1959).

Assessing Change, or Movement Toward a Goal

Because historical and evaluative knowledge concern a change taking place in complex phenomena at a certain time, in order for efforts to achieve meaningful results, it must be remembered that their context---the human mind---has sense only insofar as a goal is provided---or a course toward a specific outcome, whatever it may be. This is significant. How else can an evaluation of a given situation or process have any sense but through its assessment and interpretation of the movement (real or perceived) toward a goal or goals and the resulting changes occurring within the context of the perceptions, desires, hopes, and comprehension of those in the situation or those affected by it? In both history and evaluation, the selection of outcome, of what is taken as the close, determines

the selection and organization of subject matter with critical control being exercised with respect to the authenticity of evidential data. Truly, the desired outcome chosen, in any circumstance, determines the very selection and organization of the situation's components which are expected to move the operation closer to the goal. To evaluate effectively and with credibility, critical control for accuracy, relevancy, and significance must be exercised in gathering and, especially, in interpreting data.

The Temporal Perspective

Thus, both professionals deal with the dynamic or genetic (the becoming) as well as the static (the being or the become), and they aim at being interpretative and descriptive. For the historian this means explaining the why and how of things that happened and were interrelated, as well as what happened, when and where, and who *took* part. For the evaluator it means all of the above and in the present tense, as well. In other words, we can bridge the two professions with the temporal perspective:

The Need for Critical Sensitivity

Yet, for all the parallels, between history and qualitative evaluation, in terms of data, data collection, and data treatment processes, what is an exceptionally significant parallel is that between the cognitive abilities required of a professional historian and an evaluator. For credible, relevant and scholarly work in historical or evaluative analysis to occur, the respective professional must possess certain "gifts" or skills which relate to *practice* more directly than to theory in his field. A man who lacks common intelligence can be a physicist of genius, but not even a mediocre historian (Berlin, 1966) or professional evaluator.

> "It is only when the results of historical research have been sifted, analyzed, organized and interpreted by a mind thoroughly equipped with an understanding of human nature and social institutions that historical research may be said to have contributed in any fundamental way to permanently valuable historical writing.
>
> The real historian is not the plodder or compiling clerk who gathers the facts from a diversity of sources. He is, rather, the one who takes this raw material, evaluates it, and organizes it in such a fashion as to illuminate our minds with respect to the nature of the past and the manner in which the past has produced the present. It requires a far higher order of mind to produce historical synthesis, than to carry on historical research, and it is for this reason that there are many research students and few real historians." (Barnes, 1963)
>
> Evaluators and historians need a capacity for integration, for perceiving qualitative similarities and differences. These professionals need a sense of the unique fashion

in which various factors combine in the particular concrete situation (i.e. the perception of a social "Gestalt", as Berlin calls it. Wedgwood(1960) remarks, appropriately: "All facts are not born free and equal...there is a hierarchy of facts. To arrange them rightly, to distinguish the important from the trivial, to see their bearing one upon another, requires a skill which is very comparable to that of the painter giving significant form to the objects before him, judging the values of light and shade, or the spatial dispositions of shape and colour." Professional evaluators and historians need a capacity of association, an ability to perceive the relation of parts to wholes. They need the critical sensitivity to be able to perceive the many possible relations of the links that connect individuals seen and appreciated *as individuals* and not primarily as instances of types or laws. For a poetic manner of describing this ability, we can borrow Hegel's description as perceiving the relation of "particular sounds or colours to the many possible tunes or pictures into which they might enter (cited in Berlin, 1966, P.49)." Underlying such capacities is the ability, rooted in competence and professional familiarity with such a context, to perceive the *subtleties* present in the given situation; for, a reason is not just something to be cited...on the basis of evidence... it must be seen to be a reason, as well. (Martin, 1977).

Max Weber (1949) points out the *quality* of an event...is not something which it possesses "objectively", but that it is conditioned by the orientation of our cognitive interest, as it arises from the specific cultural significance which we attribute to the particular event in a given case. Thus, since we must ultimately rely on the significance the historian or evaluator gives to a situation being studied, we must require that, along with scholarly and professional knowledge of the contexts related to his respective field, he

possesses and develops the capacities for integration and association.

Critical Judgment and Competence

Both in historical writing and evaluation it is competence in the use of critical judgment that is decisive. A problem to be faced in both fields (one alluded to by Isaiah Berlin in his discussion of the "stream of history", "objective patterns", and "Weltanschauung") is that, at a given time, certain conceptions are so uppermost, in the culture of a particular period or group in a particular setting, that their application in constructing the events of the past (or those of the present) seems to be justified by "facts" found in a ready-made past or present. This view puts the cart before the horse. Justification if it is had, proceeds from the verification which the conceptions employed receive in the present (citing Dewey in Meyerhoff, 1959). We are reminded of Barzun's (1970) notion that judgments can be objective if they are made by testing, in all ways possible, one's subjective impressions; that in history-writing objective judgments are made every time a capable mind attends to the evidence before it. Thus, it is competence, and not majority opinion, that is decisive: even though a highly satisfactory pattern has been worked out by the historian or the evaluator, he must be willing to let it go if "facts" come to light which seriously modify it. The scholar's professional sense of form should never be so strong that he cannot modify the shape of his material when new evidence suggests that he do so. (Adapted from Wedgwood, 1960).

Key Process Questions

In historiography or professional evaluation, to determine the degree or quality of professional objectivity, the proper question is not, "Is this

an opinion?" but "What are the bases for this opinion?" Proper followup questions include: "What evidence was sought?" "What evidence was found?" "What was selected for use?" "What was actually used and for what purpose?" Because both the historian and evaluator have to deal with approximations, it is important to make sure they are as exact as possible; that they have approximated, as closely as they are able, the highest potential of professional objectivity. Therefore, it is correct to demand that both professionals be thorough, fair, and impartial in seeking and using *all* the evidence, not just evidence which agrees with their preconceptions or supports their hypotheses; it is correct to demand their professional use of critical judgment---that they develop and use a "technical" understanding of the particular cultural and motivational meaning systems involved in a given observation; it is *not* correct to demand that they be unpersons with no standards or judgments.

Conclusion

When asking "Can history or evaluation be objective?", even in the most ordinary, non-philosophical contexts, we must define and qualify such objectivity within the context of the human experience. There is no ultimate or timeless sense in which one can ask of any professional historical or evaluative account, "Is this what *really* happened?" for historical and qualitative entities are *necessarily* rooted in evaluative ideas and circumscribed by cultural aspects, both past and present. Thus, in examining the issue of objectivity in history and its implications for evaluation, the question is not whether history or evaluation be a "science". It is: "Upon what grounds are some judgments about a course of events more entitled to credence than certain other ones?"

The author posits that "objectivity" can exist in

history and evaluation, not as presenting Ultimate Truth, but within a professional context, and that the decisive factor in achieving such "objectivity" in history and evaluation is the investigator's professional competence in the use of critical judgment---his development of a "technical" understanding and, ultimately, his interpretation of phenomena.

REFERENCES CITED

[1]As quoted in Melville Dalton, "Preconceptions and Methods in *Men Who Manage,*" in Phillip Hammond (Ed.), *Sociologists at Work*, (Garden City: Doubleday, 1964).

Barnes, Henry Elmer. *A History of Historical Writing.* N. Y. : Dover Publications, 1963 (Second Revised Edition)

Barzun, J. M. and Graff, H. F. *The Modern Researcher.* N.Y.: Harcourt, Brace, and World, Inc., 1970

Berlin, Isaiah. *Philosophical Analysis and History.* N.J.: Prentice-Hall, 1966

Martin, Rex. *Historical Explanation.* Ithaca: Cornell University Press, 1977

Meyerhoff, H. (Ed.) *The Philosophy of History in Our Time.* N.Y.: Doubleday, 1959

Stern, Fritz. *The Varieties of History.* N.Y.: World Publishing Co., 1956

Weber, Max *The Methodology of the Social Sciences.* N. Y.: The Free Press, 1949.

Wedgwood, C.V. *Truth and Opinion.* N.Y.: Macmillan Co., 1960

NEW PARADIGMS FOR OLD

John W. Ratcliffe

Precis: John Ratcliffe of the School of Public Health at the University of California, Berkeley is an evaluation theoretician of social action programs. In this treatise, he "takes on" the dichotomy espoused by practicing evaluators between quantitative and qualitative paradigms and in a scholarly treatment shows the origin and fallacy of this point of view. He offers some useful methodological guidelines that should be useful to evaluators, publishers of books on evaluation that address solely or primarily quantitative issues, and educational psychologists in schools, colleges, and universities who call themselves evaluators but basically are measurement and testing experts.

In this essay I pursue three objectives. The first is to demonstrate that the traditional methodological argument that quantitative and qualitative research methods are dichotomous, polar opposites, or even separable, is viable neither from an empirical nor a philosophical perspective. The second objective follows from the first; it is to examine the implications for evaluative research that flow from the fact that all research and all data are fundamentally normative, or value-laden,

in nature. The third and final objective is to suggest some contemporary guidelines designed to assist those involved in qualitative evaluation research.

Recent Revolutions in Science.

Those familiar with the history and philosophy of science are aware that scientific progress occurs not in incremental steps but, instead, through revolutionary revisions of past scientific conceptualizations, called paradigm[1] shifts (Kuhn, 1970). Three such revolutions in scientific thought have occurred that have serious and far-reaching implications for evaluation research.

The Qualitative-Quantitative Split.

The first of these paradigm shifts is grounded in physics. The "old" physics of Newton, which has provided the basis for the ways in which most of us view the world, i.e., in mechanistic, cause-effect terms (world as the Great Machine) has been demonstrated to be invalid--that is, not demonstrable empirically and not useful in explaining the "behavior" of physical phenomena. The "new" physics of Einstein, Bohr, Heisenberg, and others tells us that the world in fact works in ways very different from those postulated by Newton--indeed, in ways that challenge at a fundamental level many of the concepts derived from the physical sciences that guide our everyday perceptions and behavior, such as the belief that the world is ordered by "natural laws," or that physics or mathematics are "exact" sciences.

Specifically, advanced research in physics has demonstrated that no meaningful distinction can be made between qualitative and quantitative approaches to the practice of science. The new physics has shown the distinction between subjective and objective to be illusory. For example, Albert

Einstein (1938) said, "Physical concepts are the free creations of the human mind, and are not, however it may seem, uniquely determined by the external world." And Bertrand Russell, co-author of *Principia Mathematica*, one of the great classics in the field of mathematics, said, "Mathematics may be defined as the subject in which we never know what we are talking about, nor whether what we are saying is true" (in Heisenberg, 1974). Indeed, the point of these recent advances in physics is that both common-sense and scientific knowledge depend at some level on qualitative knowing, for *all* such knowledge is fundamentally assumptive in nature. What the new physics has demonstrated empirically is that "facts" are determined by the theories and methods indeed *create* the facts; and that theories, in turn, are grounded in and derived from the basis philosophical (ideological) assumptions regarding the nature of the individual, society, science, and change that inform the worldviews of their formulators. At least, these are the considered conclusions drawn by many modern philosophers of science.

A test situation or survey research questionnaire provides a good example. The instructions given, whether verbal or written, require a shared language and shared motives, at least for the period of the test or interview. The responses to the test or questions necessarily involve remembered experiences, or recall which is continually subject to restructuring because of the influences of intervening experiences and the current setting. The point is that the coding and recording of responses become quantifiable properties only at the end of a *qualitative* judgmental process. And, again, it is not the meanings of the individual "yes" and "no" responses that are presented by the researcher as an end product. Instead, these data are ordered or qualified (usually aggregated), in ways the researcher deems appropriate, and the

researcher draws conclusions from the *pattern* of responses, the *relationship* between the aggregated "yes" and "no" responses. This meaning is inevitably brought to the data by the researcher, and is therefore a qualitative judgment (Ratcliffe, 1976). Data do not speak for themselves; there is always an interpreter, or a translator. Thus, any number on a computer printout, any number in an equation, or any punch on an IBM card can be traced back to a qualitative source: to the normative decisions of the researcher to follow a particular theory, to the selection of particular methods of data collection and analysis, to the interpretation of the results, and to the recommendations that follow. In short, "...much of what we think of experimental measures, recorded on the occasion of pretest and posttest for both experimental and control groups, are in fact quantifications of subjective judgments" (Campbell, 1974, P. 17). Indeed, as Bennis (1968) has asserted,

> "The reason we require 'control groups' in experimental science is that the processes presumably go on in the famous 'black box,' so we cannot observe the significant middle state of 'throughput." We can only ascertain the input and measure the output. But where it is possible to observe the throughput--the process--then the need for the crude experimental model is by-passed".

So what we have today is an unhealthy and unnatural split between quantitative and qualitative researchers: those who, in their research endeavors use, e.g., crime rates, census data, and public economic records, forgetting or ignoring their qualitative origins and criticizing those who do qualitative research; and those who critique how numbers actually get recorded (Cicourel, 1967; Kitsuse and Cicourel, 1963), and who, in a skepticism bordering on cynicism, refuse to apply quantitative techniques even in contexts where

where their use is appropriate.

One way to remedy this unfortunate situation, of course, is to bring about a synthesis of the quantitative (thesis) and the qualitative (antithesis) that are now in dialectical opposition. To do this would require researchers to recognize/acknowledge that theory, method, and data are inevitably interdependent and, therefore, fundamentally qualitative in nature. The implication is clear and unequivocal: both qualitative and quantitative approaches are, ultimately, subjective (or intersubjective) in nature, and each approach has its particular strengths as well as weaknesses. To reject either of these approaches out of hand is deliberately and systematically to limit one's horizons to a worldview--and to a research approach--that is, at best, narrow and partial.

Value-Explicit Research.

The second revolutionary shift in scientific thinking is closely related to the first, and has to do with the conduct of scientific inquiry. Following the recognition that all theory, method, and data are value-laden came the realization that scientific research performed under the traditional paradigmatic perspective has been not an "objective" enterprise but, instead, has been an *implicitly normative* enterprise, i.e., one in which the guiding value assumptions are hidden.

Thus, the fact that scientific research has tended to focus on such readily quantifiable--and, therefore, more "objective" --variables as, e.g., efficiency and productivity is not accidental. In order to maintain the *appearance* of objectivity, the moral and ethical dimensions inevitably associated with research have been deliberately and systematically avoided.

Given that value-free research is an impossibility, the emergent scientific paradigm recognizes that the values (i.e., theory, method, data analysis, interpretations, and recommendations) must be made explicit for all to see. This new worldview recognizes that any attempt now to present any research effort or finding as "objective" is misrepresentation. Making guiding value assumptions explicit in research is only fair to the consumers of and those potentially affected by the research effort; thus, it is equivalent to "truth in lending" or "honesty in advertising."

The Systems View.

The third revolutionary shift in scientific thought has occurred in evolutionary theory. The traditional paradigm is that of Darwin's, who postulated evolutionary advances occurred as the result of slow, long-term and incremental changes at the level of the individual (random mutation) as characterizing the evolutionary process. This view is still dominant, and we see its reflection in the social realm as a) incrementalism in the policy and planning processes, and b) the emphasis in social change on adapting individuals to the organization rather than vice versa, e.g., "behavior modification" in health education.

However, Ilya Prigogine (1977; 1980; 1980) won a Nobel prize in 1977 for research demonstrating that evolutionary changes were not incremental but discontinuous (i.e., radical transformations), did not take place at the individual level, but at the system level. Indeed, Prigogine and others have shown that systemic change is the norm for biological, chemical and physical systems, and that incremental change is virtually undemonstrable. Moreover, these evolutionary researchers have also found that the survival capability of a given system is determined by its ability to undergo radical structural transformations; any attempt

to reify such systems (i.e., maintain the status quo in the face of change) has a negative survival value. The point is that change in the structural dimensions of a system has greater survival value than attempting to maintain the status quo through attempting to adapt the component parts (e.g., people) to the system as given.

The social application of this paradigm views the individual not in isolation but, instead, as a "holon" (Koestler, 1979), i.e., as a system within a system within a system. The fundamental point of systems theory is that every individual entity at one and the same time is constituted of component systems (i.e., parts), itself constitutes a system (i.e., whole), and is part of a larger system. And any individual entity (system) both affects and is affected by both the systems it subsumes and the systems that incorporate it.

It should be clear that each of the three revolutions in scientific thought has substantial and substantive implications for evaluation research. These implications are the subject of the following section.

Implications for Evaluation Research

The implications for evaluation research that flow from the recent developments in science outlined above are clear and straightforward. *The first of these is that the commitment of the researcher must be not to a particular methodological approach, be it quantitative or qualitative, but to allow the nature of the problem (i.e., well-structured, ill-structured, messy, or wicked (Ackoff, 1974), defined in its particular context, to determine the specific approach used.* To reject either quantitative or qualitative methodologies based on prejudice and/or lack of familiarity with or training in the method(s) dictated by the situated

problem is to guarantee an inadequate and ineffective evaluation. If, upon analysis, the job requires methodological skills lacking in the team, the choice is either to hire consultants with those skills or decline the job. Moreover, to base an evaluation effort on methods selected not on the basis of relevance but a) because they appear to protect you from having to make critical judgments about the program or its products and/or b)because they will appear to be more "objective" to the consumer of your effort, and therefore make the study findings and recommendations more acceptable, is unethical. One facet of this implication is that, for the classes of problems (i.e., ill-structured) with which most evaluators deal in the field of education, there is *no substitute* for the application of critical judgment and reason.

The second major implication is that the values and philosophical and ideological assumptions that inform your intent (motive or purpose in performing the evaluation) and your research approach (theory, design, methodologies) must be made explicit. Given that all scientific research is value-laden, the implication is that those guiding values must be made explicit, instead of hidden, in order not to mislead those who need (or wish) to be critical consumers of the study findings and recommendations. But there is an additional reason for making guiding values and norms explicit.

Many, perhaps most, researchers are not yet fully informed of the scientific developments described above, and of their implications for scientific research standards. That is, they are not aware that methodology is theory-dependent and theory is assumption-dependent. Most competent researchers aware, however, that data are methodology-dependent; therefore, if they disagree with your findings and conclusions, they criticize your methodology. The critical concern, then is always to make explicit the inevitable determining link

between assumptions, theory, methods of data collection analysis, and conclusions and recommendations. Once the assumptions that inform your intent and overall research design have been made explicit, then it is evident to all who would criticize your work that they are actually criticizing *not* your methodology but, instead, your guiding values and motives. If your guiding values have been made clear, if they are ethically unassailable, then any criticism must necessarily be grounded in considerations that are morally and ethically indefensible.

For example, evaluations typically focus on increasing program efficiency and productivity. The justification is that these variables lend themselves readily to quantification. Such related issue areas as who benefits and who pays (physically and mentally) for increased productivity and efficiency are ignored, supposedly because they are not readily measurable, i.e., are qualitative in nature. Yet, as we have seen, *all* data are qualitative in nature, including those pertaining to efficiency and productivity (e.g., in the definitions of the terms: efficiency for whom, the worker or the employer?). Variables related to issues of equity and social justice are not more difficult to measure (i.e., judge) than efficiency and productivity, as we shall see in the following section. *It is just that we have had less practice and experience in these areas, because of the implicit values that have been guiding evaluative research under the old, so-called "objective" paradigm.*

The third implication for evaluation research is that any adequate evaluation effort necessarily assumes a systems perspective. That is, personnel and/or programs must be viewed as system components (as holons, or systems within systems within systems), rather than as independent entities, in order to obtain a proper perspective on the impact

of forces acting upon them and generated by them. The fundamental point of the systems approach to evaluation is that parts are not wholes; parts and wholes are simply at different levels, and neither can be defined (and certainly cannot be evaluated) without reference to the other. Since they are at different levels, they therefore require different levels of analysis; but an analysis of the dynamics of their interrelationship is clearly an essential dimension of understanding either part or whole. For example, in personnel evaluation it should be recognized that individual behavior is rarely idiosyncratic, but is heavily influenced by and, in turn, influences the context within which it occurs. In a similar vein, the outcome of a program is not to be explained only in terms of its own constituent components, but must also take into account the relationship of influences between the program and the broader system of which it is a component.

The criteria, then, for a comprehensive evaluation effort are: logical and rigorous design[2], systematic and holistic treatment (including constant attention to the wider context), clarity and congruence of undergirding motives and assumptions, and a persuasive presentation of why the theory and method(s) selected were particularly suited to this instance. How these numerous requirements can be achieved (operationalized) is the subject of the next section.

Methodological Guidelines

Methodology differs from method in much the same way that strategy differs from tactic. Methodology, like strategy, is the guiding set of general *principles* that informs selection from the possible range of alternative methods of operational procedures, or tactics, through which the ultimate goals of the research project are attained. Thus methodological principles are typically general in

nature, and are context-free in that they apply to most settings. Methods, on the other hand, are usually context-specific in that they can be generalized to other settings with much less confidence. The following sets of methodological principles should be viewed as providing general guidelines. They have been selected because of their particular usefulness to those in the field of education, where ill-structured[3] (i.e., hard to define) problems are the rule rather than the exception.

Applied Ethical Analysis

This is the essential component of all evaluations, for this is procedure whereby the motives, values, and ethical considerations are explicated. This normative or value component sets the overall tone for the research effort; it is the bass cleff of the evaluation band. It usually has two dimensions. The first deals with clarifying the value assumptions of the evaluator(s), through such sets of questions as these: 1) what are your motives in undertaking this evaluation? Whose interests and purposes will be served? Yours? The funding agency's? Those who should be benefitting from the program?; 2) what philosophical and ideological assumptions will guide this evaluation effort? What are the implicit assumptions your preferred psychological and sociological theories hold regarding the nature of the individual (e.g., responsible for his/her own situation; society is responsible; holon; etc.) and of society (e.g., reflects the will of the people; imposed on the people; mutual-causal; etc.)? What are your assumptions regarding the nature of social justice and equity (e.g., utilitarian; pragmatic; Rawlsian, etc.)? If the evaluation team consists of more than one member, as is usually the case, the assumptional analysis technique described later can be usefully applied here.

The second dimension deals directly with the degree to which the program being evaluated is ethical in nature. This is delineated through such questions as the following: who is the program *intended* to serve, and who does the program *actually* serve (e.g., the staff, the managers, the funders, others)? The degree to which there exists a discrepancy between intended and actual beneficiaries is the degree to which the program is unethical or immoral. Another question to be posed is, who benefits and who pays? In this question, the costs and benefits may be tangible (e.g., profit) or intangible (e.g., psychological stress), and both direct and indirect costs and benefits must be measured.

Stakeholder Analysis

The traditional and still-dominant approach to evaluation is goal-based evaluation. In this approach, the evaluator is told what a management's view of program goals and objectives are, and the evaluation effort proceeds to assess how well those goals and objectives have been met. The major limitation of this approach is that only one perspective of the program has been given: management's, i.e., the official version. The view of others who hold stakes (interests) in the outcome of the evaluation is missing from the equation.

Stakeholder analysis is based on the view that truth is *synthetic*, i.e., the perspective of the program that is most useful in guiding an evaluation stems from comparing and contrasting as many alternate views as possible. This means, in effect, that an adequate evaluation entails exploring in some depth the views of each stakeholder group involved (e.g., funders, management, staff, direct and indirect beneficiaries, etc.) regarding program objectives, function, value, merit, who benefits and who pays, etc. If alternate views surface, as they commonly do, data relevant to each

are collected in order to compare the various data bases for congruence of fit.

Assumptional Analysis

This methodological approach is particularly useful in formative evaluation; it is also quite useful in metaevaluation (see the Key Evaluation Checklist below). The focus is the underlying value system of the program/operation being planned, through applying methods that will bring to the surface the values implicit in decisions taken regarding policy issues, means or tactics, near-and medium-term objectives, and long range goals. It is basically a dialectical approach, where a single data set is subjected to conflicting interpretations. Again the truth is considered to be synthetic, i.e., it will emerge through the process of argument.

Key Evaluation Checklist

Evaluation is a complex process. It is usually necessary to explore so many important dimensions that some can be overlooked or "lost" without a checklist of some kind. The following checklist, adapted from Scriven (1980) and presented in very brief summary form, outlines those items considered to be dimensions essential to most evaluations. The most important dimensions are asterisked.

1. Description. What is to be evaluated? What are its component parts? Of what system or systems is it a part?
*2. Function. What are the purposes, goals, and objectives of the program? Distinguish between *alleged* or *intended* function and *actual* function through stakeholder analysis (see above).
*3. Applied Ethical Analysis (see above).
4. Resources. Analyze the resources (e.g., money, expertise, technology, flexibility considerations) available for use by a) the

program and b) the evaluators.

*5. Delivery system. How does the program get to the consumer/beneficiary? How is it maintained? How is it implemented, updated, monitored? Who does all this?

*6. Needs Assessment. What are the different needs and wants of the different stakeholder groups? Are they being met by the program?

7. Standards. Are there "standards-in-use" that will help to assess the program vis-a-vis others of similar nature?

8. Process. What constraints/costs/benefits apply to the normal *operation* of the program? E.g., how well does the operation match the program description?

*9. Outcomes. What effects (intended *and* unintended) are produced by the program?

10. Costs. Dollar vs. psychological; initial vs. repeated; direct/indirect vs. immediate/delayed, etc.

*11. Comparisons. The program should always be evaluated with regard to its "opportunity costs," i.e., are there alternative options (recognized and unrecognized) that might more effectively achieve the stated goals?

*12. Significance. A *synthesis* of all the above. This is the most important, and most difficult and complex, dimension of the evaluation process. This is where the exercise of critical human judgment is most heavily brought to bear; it is the responsibility not of management or any of the stakeholders, but of the evaluator(s).

13. Recommendations. May or may not be requested, and may or may not follow from the evaluation.

14. Report. Logistics like target group(s), vocabulary, length, private/public, screening, circulation of drafts, etc.

*15. Metaevaluation. The evaluation must itself be evaluated, preferably prior to a) implementation, and b) final report dissemination.

External evaluation is desirable, but building the metaevaluation into the checklist allows for self-criticism, which is preferable in the long run. One of the key questions in the metaevaluation is, whose interest will the evaluation serve?

Conclusion

Discoveries in the natural sciences have always exerted heavy influences on our ways of thinking about and acting in the world of which we are a part. For example, the discoveries of Copernicus turned our view of the world from the hub of the universe to one in which it is but a speck in a lesser solar system in a minor galaxy on the periphery on the universe. In this essay I have discussed in summary form some recent revolutions in scientific thought. While less sweeping in philosophical scope than the Copernical revolution, these paradigm shifts have significant and substantive implications for the conduct of evaluative inquiry. These implications were briefly outlined, and some methodological guidelines for evaluation were presented that are more congruent than the traditional approaches with what is known today about the way the world is constructed.

FOOTNOTES

[1] A paradigm is defined here as a world view, a way of ordering and simplifying the perceptual world's stunning complexity by making certain fundamental assumptions about the nature of the individual of Society, of Science, and of Change. Paradigms are normative; they determine what the practitioner views as important and unimportant, reasonable and unreasonable, legitimate and illegitimate, possible and impossible, and what to attend to and what to ignore. In learning a paradigm, one acquires theory, method, and standards together, usually in an inextricable mixture. Moreover, through the theories they embody, paradigms prove to be

constitutive of all normal scientific activity, including underlying assumptions made, problem definition, areas of investigation, questions posed and, particularly, data interpretation and policy recommendations made at the conclusion of the research process (Kuhn, 1970). Thus, all theories as well as the strategies derived from them are, ultimately, paradigm-based.

[2] When a quality, or property, is quantified (e.g., grouped, aggregated, inferred, etc.) it can be done appropriately or inappropriately. That is, it can be applied to the wrong context, the data to which it applies may not meet the assumptions underlying the method, etc. Similarly, when a quality is qualified (e.g., ordered, categorized, classified, processed through description, synthesis, etc.) it can be done appropriately or inappropriately. It may be done in the wrong context, it may not follow from stated guiding assumptions, etc. The point is that qualitative is *not* simply another name for sloppy research!

[3] Ill-structured problems are complex mixes of highly interdependent problems that are exceedingly difficult to formulate, let alone resolve independently of one another. Organizational or programmatic ill-structured problems are those in which more than one person is involved in their formulation, solution, implementation, and evaluation.

REFERENCES CITED

Ackoff, R.L. *Redesigning the Future* New York: Wiley, 1974

Bennis, W.G. "The Case Study." *Journal of Applied Behavioral Science.* 4:26 (1968), 227-231.

Campbell, D. T. "Qualitative Knowing in Action Research." Kurt Lewin Award Address, Society for the Psychological Study of Social Issues,

Meeting of the American Psychological Assn., New Orleans, September 1, 1974, P.17

Cicourel, A. V. *The Social Organization of Juvenile Justice.* New York: Crane-Russak, 1977

Einstein, A., and Infield, L. *The Evolution of Physics.* New York: Simon and Schuster, 1938, P. 31.

Garfinkel, H. *Studies in Ethomethodology.* New York: Prentice-Hall, 1967.

Heisenberg, W. *Across the Frontiers.* New York: Harper and Row, 1974.

Kitsue, J. L., and Cicourel, A. V. "The Use of Official Statistics." *Social Problems,* 11:2 (1963), 131-139.

Koestler, A. Janus: *A Summing Up.* New York: Viking, 1979.

Kuhn, T. *The Structure of Scientific Revolutions* Chicago: University of Chicago Press, 1970.

Ratcliffe, J. W. "Analyst Biases in KAP Surveys: A Cross-Cultural Comparison." *Studies in Family Planning,* 7:11 (November, 1976) 322-330.

Scriven, M. *Evaluation Thesaurus.* Inverness, CA: Edgepress, 1980 (in press).

TO SEE WHAT IS HAPPENING

Mark St. John

Precis: At the University of California, Berkeley, there is a unique organization of scientists and educators known as the Graduate Group in Science and Mathematics Education (SESAME). Mark St. John has been a member of that group for the past five years. Despite his close association with faculty members drawn mostly from the hard sciences, he is interested in qualitative research. In a creative and persuasive essay, he articulates and demonstrates the validity of the naturalistic, illuminative approach to evaluation.

An ancient fable tells the story of a young man who lived in an isolated village along a river in China. Feeling restless in the small town he sought adventure, and one day had the idea of exploring new lands by floating down the river on a raft. He walked through the village seeking information about the river, but few people had ever gone more than a few miles from their homes. Finally, he found some elders who could remember stories they had heard about the river, and he also found some drawings in an old book. From these, he was able to piece together a map of the river. He built a strong raft, gathered some food, and set out on

his journey.

After a few miles the river took a sharp right turn, as indeed the map said it would. In a mile the river made a fork around a small isle, as the map indicated. His confidence grew, and he felt secure and happy. But in two miles, where the map indicated a right turn, the river went straight. Next, where the map showed forests, there were lush meadows. As he continued, the young man saw more and more discrepancies, and he grew distraught with the feeling that he was now on the wrong river. His despair increased until, in a flash, he saw the source of his problem. He threw the map overboard; and for the first time he looked at the countryside as it was, and negotiated the river as it flowed.

As educators do we see clearly what is actually happening in our classes, or do we primarily see our own maps--those models and images we have gained from our experiences and training? Much of educational research, by validating and testing ideas, attempts to build maps of where we should go. Similarly, it could be said that education evaluation primarily measures progress along some of the routes we have mapped out. While it is useful to build visions of where we want to go, it is equally important to know where we are. Is our present description of the educational process sufficiently accurate, complete, and powerful to enable us to move from where we are?

In the past five years there has been a growing school of thought in the community of educational researchers and evaluators, that it may be worthwhile to expend considerable energy in simply trying to see, and imaginatively portray, what is actually happening in our educational institutions. The purpose for such research is to expand our understanding of what we are doing and, through creative redescription, to allow insights to occur which will guide us to more effective actions.

Evolving is a general approach, that includes many research and evaluation models, which has the following characteristics:

(1) The approach is naturalistic in the sense that it studies educational programs in their natural context without experimental manipulation. Moreover, the approach is naturalistic in that it attempts to discover the nature of the situation it studies--that is, those inherent, inseparable properties which are essential to the situation's fundamental character.

(2) The approach is qualitative in that it attempts to discover the critical qualities and dimensions which are relevant (as opposed to measuring quantitatively along predetermined dimensions).

(3) The approach is illuminative in that it attempts to shed light on, to make clear, and to elucidate the situation it examines.

(4) The approach is holistic in that it attempts to see the whole and complete picture and to synthesize rather than to analyze.

(5) The approach focuses on process in that it attempts to see continuous and regular actions, or larger patterns of action, which are taking place.

The intention of this approach (henceforth, called the "qualitative approach") is to see the facts of what's going on or being carried on. This seeing is different than the process one goes through in measuring. According to Stake (1977):

> "It is natural to see. It is natural to measure. Seeing and measuring are not the same act.
>
> "But they are even more different than we

supposed. The common notion is that when one measures, one sees the same thing, but sees its amount. As if one was seeing through glasses having graduated-scaled markings on them. Measurement glasses, however, do much more than scale the view. Much more difference there is between seeing and measuring.

"There is a transformation from experiential perception to representational perception. The observor switches from actor to director. He/she gives up the direct impression of the thing, perceiving it no longer as another being, the whole object, a member of the physical populace, and perceives it then as bearer of properties, or even merely as an array of characteristics. This is no small transformation...

"The way most of my researching colleagues want to see the world is through the properties of things. The way most of my teaching colleagues want to see it is to see things as things.

Putting on glasses that focus on properties, scales, and amounts changes the perception. Perhaps only a little, as sunglasses do; perhaps a lot, as reversal prisms do...

"Measurement is not just holding a ruler to what we see, but seeing something to hold the ruler to."

The essence of the qualitative approach is this "seeing" - an observation which lets the situation reveal its properties and qualities, rather than forcing perception through the researchers preconstructed template of ideas. More specifically, using the qualitative approach one observes the people in the setting, their individual behavior and their interactions; one talks with the participants in an attempt to learn the understandings and

meanings they ha e created; and one studies data, records and documents, for what they contain explicitly, and for what they show about the point of view from which they were created. Much as an anthropologist throws himself into a new culture, the qualitative researcher must throw himself into the educational world he studies.

Most importantly, the researcher cannot filter what he sees through his old system. Like learning a foreign language, he must use the "speak it all the time" method, rather than the translation method. He "must internalize the ideas so that their connections are firmly engraved in his memory and his reaction, and can be reproduced at will...this process must be kept free of external interference. For example, the researcher must not try to get a better hold on the ideas by likening them to ideas he already knows, or finds more comprehensible or more precise. On no account must he attempt a 'logical reconstruction'. Such a procedure would tie him to the known, or to what is preferred by certain groups, and would forever prevent him from *grasping* the unknown ideology he is examining." (Feyeraband, 1979, P. 250)

Thus is born the idea of the participant-observer. Rather than see only what he can clearly express, the qualitative researcher must first learn and incorporate the situation in his own mind, and then struggle to express, to translate what he now *knows* into the language which his readers are familiar with. In operation here is an incommensurability principle which disallows the researcher (or anyone else) from seeing, knowing, experienceing the situation being studied and at the same time describing it in the precise, clear, analytical terms of an external language.

What is the point of this qualitative research? How does it pay off? Who uses it and how? To investigate such questions we will first examine how this

approach may be used to provide feedback to an individual educational program--the case of qualitative evaluation. Next we will consider how the qualitative approach can serve a wider audience-the case of qualitative research.

Qualitative Evaluation

To see how qualitative evaluation can serve an on-going program, three examples are presented in this section. All three evaluation examples come from my own experiences as an evaluator in the past few years, and they illustrate some of the more general functions which qualitative evaluation may serve.

(A) The Physics Department at the University of California at Berkeley has a small-scale program designed to train their graduate teaching assistants. Having volunteered to run the program, I was faced with the problem of finding ways of interacting with the teaching assistants so that they could improve their own teaching. One possibility was to present to them some general principles of good teaching and some proven teaching techniques (as well as some common pitfalls). While the teaching assistants could understand these ideas, and even agree with them, such ideas did not ultimately empower them to alter their own behavior (just as knowing the benefits of non-smoking often does not empower the smoker to quit).

Thus, instead of illustrating how teaching should or shouldn't be, I adopted the basic approach which underlies qualitative evaluation--we began with an attempt to see what was happening. The teaching assistants took turns video-taping each other in their classes and each week we would spend several hours simply watching the video tapes. We did not have a pre-established set of behaviors to look for, nor did we try to determine what was good or bad about the teaching. Rather, we looked for patterns in the TA's and student's behavior. We also tried

to see (feel) the TA's mental and emotional experience as it occurred in front of us on the screen.

In watching their own tapes the TA's could see what they were actually doing. This is a non-trivial feat. At the time of their teaching the TA's were much too involved in presenting material and managing the class to notice patterns in what was happening. Also barring insight into their own behavior and experience was a fairly solid self-description of their own teaching behavior and the reasons for it being the way it was. Observation of the video tapes, without judgment and therefore defensiveness, allowed the TA's to re-experience what they were doing, but this time with awareness. At times we would even stop the tapes and have the TA's stand up and repeat the same behavior (without any effort to change it). In this way they could actually see their own behavior patterns and sense them as they occurred, and they could pierce through their own self-descriptions. It was only at this point that the TA's had a direct enough access to their own behavior that change might happen. Thus, in this class the qualitative approach--the seeing without measurement, comparison or judgment--allowed the process of change to begin on actual and not just a theoretical level.

(B) In the Winter of 1979 the American Association of Physics Teachers held its annual meeting in New York City. I and another researcher carried out an illuminative study of this meeting with the purpose of writing up a portrayal of the meeting and presenting it to all the members of the AAPT. We observed the participants in the meetings and in the hallways; we sat and ate and talked with them; we asked them to fill out questionnaires before, during and after the conclusion of the meeting. From our experiences and our data we wrote a report, published in the American Journal of Physics (St. John and Cid-Campos, 1979), which presented the

participants' goals, perspectives, and experiences.

We isolated and elaborated on a few issues that we thought were critical in the minds of the members and the functioning of the meeting. By summarizing the various points of views and concerns, we were able to capture a sense of the meeting as a whole that was unavailable to any one participant. After considerable debate several changes were made in the next year's meeting. If these changes are effective, it is perhaps because they came out of a larger and clearer perspective of the whole situation. In this case our report sought to fulfill the goal of illuminative evaluation that was well described by Parlett (77):

> "The task (of illuminative evaluation) is to provide a comprehensive understanding of the complex reality (or realities) surrounding the program: in short, to illuminate. In his report, therefore, the evaluator aims to sharpen discussion, disentangle complexity, isolate the significant from the trivial, and to raise the level of sophistication of debate." (p.21)

(C) Physics 5-I is an introductory course designed for science and engineering majors and uses the PSI (Personalized System of Instruction), or Keller Plan format. According to this format students use well-designed materials, work at their own pace, and master the course material in a stepwise fashion, passing successive module tests before they are allowed to proceed. Another researcher and I studied this course for two quarters carefully examining the distinction between the course as designed and the actual events that took place as the course unfolded (Cid-Campos and St. John, 1980).

For example, the instructor designed the study materials to be used according to a certain plan--students should begin with the module objectives, read the required readings in the text, and at the

end work practice problems to test their understanding. In contrast, many students, to maximize their time efficiency, began with the problems and interacted with the text only as much as was necessary to solve the problems. As one student said:

> "After skim reading, I see what they find important by doing a problem. Then, with that problem I'll go to the book, and I'll look up how the problem pertains to the book, and how the book pertains to the problem."

There is clearly a mismatch between the instructor's ideas of how the students are (should be) studying and the actual practices the students develop. As a result of this mismatch the instructor spends a great deal of time and attention writing elaborate text prose (which isn't read), and very little time on choosing good problems and writing useful solutions (which are heavily used).

When we examined the concept of mastery, we found the grading of the students by the tutors on the module tests depended on many factors other than the students' performance. We saw patterns of interaction between the students and the tutors that depended on their social connections. We discovered that the tutors thought of themselves primarily as teachers (and not as graders). They preferred to use these interactions as a teaching forum rather than strictly adhere to the grading procedure. We saw that the students and tutors conspired (in an implicit way) to avoid confrontation and to allow the students to proceed through the course on schedule. As one TA said:

> "You are trying to get people through this class while still learning as much as possible, and maintaining your sanity by not nitpicking every single point. Sometimes you get too lax, and sometimes, well,...you just pass the person."

Our report was disillusioning to the instructor in the true sense of the word. He was planning, operating, and thinking about the course using the Keller Plan design as a mental template. He thus did not see the larger perspectives and patterns of behavior that were actually taking place in the course. Our report allowed him to see where, and perhaps more importantly where not, to put his efforts in an attempt to improve the situation.

Qualitative Research

The goal of most educational research is to contribute to a body of knowledge about education and the educational process which will allow us to make unambiguous predictions. Typically, on the basis of a particular experience or observation, researchers form hypotheses (which they usually feel to be true) and design experiments in such a way that their hypotheses can be proven. If clever, they will test them using random methods over a sufficiently wide population that their own conclusions may be thought to be generally "true". Such conclusions then get added to the body of knowledge and, theoretically, become available for educators to use in their teaching and designing of materials.

Because the researcher focuses on the creation of a proven body of knowledge (which will hopefully be of use in the future), there is almost no overlap of the worlds of the educational researcher and the practitioner. There are other reasons for the lack of connection. *Educational researchers aim their papers at other researchers.* The results of their research are communicated in such a way that the major emphasis is on establishing the proof or validity of the results rather than on effective communication with the reader. Also, there is the inherent difficulty that behavior is hard to change and a general abstraction, even proven by research, is unlikely to convince people to change. More often, perhaps, practitioners use research results

to justify what they already want to do.

The qualitative approach to research attempts to capture and share generalizable experiences rather than establish proven truths. The goal of the qualitative researcher is to produce empowering communications, not to find true, immutable knowledge. This distinction between capturing and communicating experience as opposed to proving and defending formal conclusions is a very important one. Formal conclusions, to be proved and defended, must be in the realm of the observable, and therefore must be behavioral rather than experiential. Such a methodological constriction can lead to a false notion of cause and effect.

An example of this confusion can be seen in the research done on "wait-time", where it has been proven that teachers who wait after asking a question are able to raise the level of sophistication of the students' responses and subsequent discussion. Carefully done experiments lead to the conclusion that increasing wait-time *causes the* change in the level of discussion. An alternate explanation, and one that seems more consistent with personal experience, is that wait-time is a natural manifestation of an intention to communicate. That is, if a teacher asks a question with a real intention that it be considered seriously, and really does want the students to think it out for themselves, then the students will sense the intention behind the question. The teacher will naturally wait for them to consider, and the students will naturally take their time to think the question through. Thus, it is actually the communication of the intention that raises the level of debate, and the wait-time is a natural byproduct. However, since intention is not measurable, the researcher creates a variable called "wait-time" and correctly associates it with a higher level of sophistication. But then to train teachers to increase their "wait-time" behavior independent of

their intention is to miss the whole point.

This is why the qualitative approach focuses on experience itself rather than formal conclusions abstracted from experience. An eloquent spokesman for the qualitative approach, Elliot Eisner (1979), calls for the use of a methodology he calls "educational criticism" which he describes as a parallel to art and literature criticism. The role of the critic, Eisner argues, is not to discover true abstractions but rather to disclose and communicate experience.

Similarly, Bob Stake (1979) points out a way in which a researcher may help a wider audience through the sharing of experience:

> Most researchers know that they seldom can produce generalizations that will be of direct benefit in solving a particular practical problem. Instead of trying only to improve the researchers production of generalizations, an alternative is to redirect some of the research effort into producing some of the ingredients for the user's (own) naturalistic generalizations. The researcher in this role becomes the *facilitator* of generalizing rather than the producer of generalizations, more the discovery-learning teacher than the didactic teacher."

The "naturalistic generalization"he speaks of is described by David Hamilton (1979) in the following way:

> "Naturalistic generalizations are inside the head generalizations...located within the realm of private knowledge. Naturalistic generalizations are personal accounts of the external world which, as it were, are held in a non-negotiable currency."

Thus, much like a good novel allows one to have

insights into one's own life, so can qualitative research allow an educator insight into their own educational situation. Such insight is personal and involves the reconstruction of mental meanings and understandings. Through vicarious experience one is changed, and without having to remember or refer to formal conclusions, one will think and act out of a new perspective.

According to Eisner (1979), the criteria for judging the quality of a piece of qualitative research is not its correspondence with reality (which is unknowable) but rather its usefulness. Usefulness is defined as "whether it allows one to do one's work more effectively, whether it enables one to perceive the phenomenon in more complex and subtle ways, whether it expands one's intelligence in dealing with important problems." (p.214)

An example of a large-scale qualitative research project which was intended to be of use to a wide audience was the Case Studies in Science Project funded by the National Science Foundation in 1976. The goal of the project was to investigate the actual status of science and mathematics education as it was occurring in the classroom throughout the United States. Overseen by Bob Stake, qualified researchers carried out qualitative case studies at eleven different sites across the country. Researchers attended classes, talked with teachers, parents, administrators, and embedded themselves as thoroughly as they could in the local community for time spans ranging from three weeks to three months. Each of these eleven researchers wrote a case study which illuminated the situation in the district they studied, and they explicated the critical factors they saw impinging on science and math instruction. The case studies attempt to convey what it was like to be there, what the daily problems, successes, and failures looked like to the parents, teachers and students.

These case studies are useful in two different ways. First, by examining all eleven studies as a whole, one can look for common patterns and factors effecting science and mathematics education. In this way formal conslusions can be drawn in a somewhat parallel fashion as in an experimental study. For instance, in this study, it was a general finding that little science education was taking place largely because of the pressure of the "back to basics" movement and increasing teacher accountability in this area. Another formal finding, abstracted from the case studies, was that the textbook was still the main source of instruction, and that teachers were still largely using didactic recitation methods. There was little experimental work and little teaching based on an inquiry method. Such conclusions are useful (to policy makers) in that they can point to where action is needed and where it may(and may not) be effective. For instance, the case studies suggested that teacher training institutes are now probably a more effective way of improving science education than the development of a new curriculum.

A second way in which the case studies are useful is not in their conclusions but in their portrayals. They allow policy makers, administrators, educators, and researchers a chance to personally experience what it is like to be there and to see the complexity and variety of educational issues at different sites. Reading through the case studies one begins to see the simplicity of the models (maps) we have created to describe instruction and learning, and one gets the sense that the lives of these teachers and students is no less complex than one's own. Thus, the case studies may be useful in that they destroy simplistic notions and suggest that there are no easy external fixes for the problems we face. Such a finding may throw us initially into hopelessness, but ultimately it may allow us to see and to deal with the actual facts of our educational system.

Some Final Thoughts

This distinction between seeing what is happening and one's image of what is happening is very subtle and important. An artist friend of mine describes it in the following way:

> "In painting an object (say one's hand) one would like to paint what is there--that is, to paint the hand as it actually looks in its particular configuration. The knowledge of what a hand looks like (the past stored images that are associated with that word) only get in the way. As you begin to paint, you find yourself painting the picture the way you think it should look, rather than painting the actual image which is in front of you.
>
> "In painting there is an obvious test to see if you are looking at the hand as it is--you can compare the image of your hand as you see it with the image you are creating on the paper. The discrepancy between the two images tells you where the stored image took over and where there was no actual observation of the object. In this way the picture you paint (the product) serves as a mirror of your own attentiveness.
>
> "The most useful perspective to have in examining the painting (the product) is not to judge whether it is good or bad, or to compare it with others, but rather to use it as a guide to notice where the seeing of the object stopped and where the image took over. In this way the product can tell us where we are not seeing the facts."

Perhaps in the same way we can use a qualitative approach in educational research to see where we've stopped dealing with what is actual (what is happening) and where we started to deal with our image (our models, maps, and memories of how the

situation should be). We must be careful, however, not to embrace qualitative methodology too closely. The word "method" comes from two Greek words meaning "mid" plus "way". Thus, method is the "central path", and it is something which you follow. The problem with method is, that as a path to be followed, it prevents seeing what is actually happening. Thus, one's action may come out of the method, blindly in accordance with a set of rules, rather than in response to a knowledge of the facts. Eisner (1979) points out the danger of believing in our methodology: "The hunt for recipes, rules, formulas and other nostrums to solve other educational problems is a hopeless one.. we need, I believe, to recognize the contingent character of educational practice, to savor it's complexity, and not be afraid to use every artistry we can muster to deal with the problems."

REFERENCES CITED

Cid-Campos, R. and St. John, M. (1980) The Inner Workings of PSI: A Case Study of an Introductory College Physics Course. Submitted to the American Journal of Physics, July 1980.

Eisner, E. W. (1979) *The Educational Imagination: On the design and evaluation of school programs.* MacMillan, New York.

Feyerabend, P. (1979) *Against Method.* Verso, London, England.

Hamilton, D. (1979) Generalization in the Educational Sciences: problems and purposes. Unpublished paper. Glasgow, Scotland.

Parlett, M. and Hamilton, D. (1977) Evaluation as illumination: a new approach to the study of innovatory programs. In Hamilton et al (1977). *Beyond the Numbers Game,* McCutchan, Berkeley, California.

Stake, R. and Easley, J. (1978) Case Studies in Science Education. A series of reports available from CIRCE, University of Illinois, Urbana, Champaign, Illinois.

Stake, R. (1980) Generalizations. A paper presented on April 9, 1980, at the Annual Meeting of the American Educational Research Association, Boston, Massachusetts.

Stake, R. (1977) On Seeing and Measuring. In a newsletter issued by the Case Studies in Science Education Project. CIRCE, University of Illinois, Urbana, Champaign, Illinois.

St. John, M. and Cid-Campos, R. (1979) A Look at the 1979 AAPT Winter Meeting. American Journal of Physics. 47 (December 1979) 1024-1030.

CAN EVALUATION PROVIDE USEFUL INFORMATION?

Rebecca Zwick and Leonard A. Marascuilo

Precis: Rebecca, Zwick, a doctoral student in Measurement and Evaluation Research and Statistics at UC, Berkeley and her mentor and well-known statistician, Professor Leonard Marascuilo, have dissected the merits of quantitative vs. qualitative education and present a persuasive argument in favor of the former over the latter. Their strawman is a qualitative evaluation study of their choice and they show how it would have been improved with their own quantitative approach.

1. Why Has Quantitative Evaluation Been Viewed As Unsuccessful?

During the past few years, evaluators have become increasingly critical of program evaluations based on the quantitative and statistical models customarily used by psychologists. The reason for this criticism is that evaluation after evaluation has shown that educational interventions have not been successful in meeting program goals (e.g. see Filstead, 1979). That quantitative evaluations often have been unable to demonstrate positive results cannot be denied, but to argue that such methods therefore are invalid and should be

abandoned and replaced by qualitative methods is questionable and might be likened to throwing out the baby with the bathwater.

Why have quantitative evaluations often failed to detect program effects? One reason is that some evaluators have insisted on utilizing standardized tests to assess program outcomes, a practice which is rarely appropriate. Standardized tests usually do not contain items which are relevant to the content of a program to be evaluated; they will tend to be insensitive to the specific changes which are expected to occur. In addition, the samples on which evaluations are based often are totally dissimilar to the population used for norming a standardized test. Too often, evaluations have focused on questions which were only of peripheral interest, and thus failed to provide maximally useful information to decision makers. Another reason for the failure of quantitative methods is that program effects, if they do indeed exist, may be small in magnitude. Few evaluators have addressed the question of how large an outcome must be before it can be considered significant. As is well known, statistical significance does not necessarily imply educational significance.

Based on these shortcomings some evaluators have chosen to conclude that any technique which smacks of the experimental method must be eschewed and that the fallibility of traditional methodologies somehow establishes the superiority of alternative evaluation strategies. The authors believe that evaluators should be skilled in a wide spectrum of techniques, choosing those methods which can best provide the type of information being sought. In general, qualitative methods are to be preferred when process variables such as teaching style, program implementation, the classroom interactions are being evaluated, whereas quantitative methods are appropriate for the assessment of program outcomes such as school achievement, attitude change

and attendance.

2. What Other Objections Have Been Raised Against Quantitative Evaluation?

Some evaluators are opposed to the use of quantitative methods, not because of their perceived insensitivity to treatment effects, but because they are viewed as philosophically objectionable. A recurring theme in the qualitative evaluation literature is that quantitative evaluations reduce people to numbers, whereas qualitative evaluations are more humanistic. Many seem to believe that quantitative methods destroy the hearts and souls of the individuals being assessed so that their human identities exist merely as holes on keypunch cards that become the food of the computer; people become statistics and lose their human qualities. Parlett and Hamilton (1977) liken conventional evaluation studies to research done in the areas of agriculture or botany: "Students--rather like plant crops-- are given pretests (the seedlings are weighed and measured) and then submitted to different experiences....Subsequently,....their attainment (growth or yield) is measured to indicate the relative efficiency of the methods (fertilizers) used". (p. 7)

Parlett and Hamilton, as well as Guba (1978) believe that naturalistic approaches, unlike conventional quantitative methods, allow the evaluator to become truly informed about the learning milieu and program participants. The prevailing idea is that while the quantitative evaluator is in his modern-day ivory tower, punching buttons on a computer terminal, the qualitative investigator is out in the field, "getting close to the data." It is easy to empathize with Fienberg (1977), a statistician, when he says, "I find it astonishing that getting close to the data can be thought of as an attribute of only the ethnographic approach. Perhaps statisticians are in fact ethnographic

researchers in disguise, for the good statistician working on a project as a collaborator tries to learn all about the data before designing an experiment or planning a sample survey". (p. 51). It is equally difficult to understand the basis for the mysterious belief that the use of numerical techniques violates the human qualities of the individuals being assessed.

3. How Do Qualitative and Quantitative Evaluations Differ?

Traditionally, quantitative evaluation suggests to most evaluators and consumers a model in which a program evaluator has identified program goals and objectives, has operationally defined independent and dependent variables, has developed questionnaires, tests, and rating scales, has subjected them to a pilot testing and examined them for internal consistency, reliability, and validity, administered them to experimentally sound control groups, coded the responses for computer analysis, statistically tested the relevant hypotheses, made examinations of probability levels, and produced a written report.

The prototypic qualitative evaluator, on the other hand, is largely an ethnographer who becomes immersed in the operation of the program, makes careful and detailed observations upon the teachers, students, administrators, and parents as they act and interact with one another, and conducts in-depth interviews with participants, generating a rich source of data through the use of open-ended questions concerning the interviewee's feelings, opinions, and insights about the program.

For the most part, these characterizations represent a generally accepted, if somewhat stereotypic, view of the two models. A review of the literature reveals, however, that there is a lack of consensus concerning the definitions of qualitative and

quantitative evaluation.

Guba (1978) attempts to differentiate what he calls naturalistic inquiry from conventional inquiry. These forms of inquiry appear to correspond roughly to qualitative and quantitative evaluation, respectively. A major difference, according to Guba, is that conventional inquiries are designed to examine scientific facts and their relationship to one another, whereas naturalistic inquiries are done for the purpose of understanding social phenomena.

Conventional inquiry, says Guba, involves identification of independent and dependent variables, experimental and statistical control, random selection or assignment of subjects, and decision-making based on probabilistic considerations. Naturalistic inquiry involves immersion into the setting and permitting impressions to emerge. The problem with this is that two persons immersed in the same world, or viewing the same microcosm generated by a program, might find that their impressions differ, perhaps by as much as 180 degrees. Guba, on the other hand, might argue that the risks of such divergent interpretations would be minimal, provided that the investigator used triangulation procedures to verify the findings.

Guba also sees the conventional investigator as a *reductionist*, or one who seeks answers only to the questions formulated prior to the inception of a program. On the other hand, the naturalistic investigator is an *expansionist*, or one who will let the program suggest questions to be answered. According to Guba, the conventional evaluator wants to test hypotheses; the naturalistic evaluator wants to gain insights about the complexities and inner workings of a program. In our view, any competent evaluator would want to do both.

Guba sees quantitative evaluation as locked in step;

once a design is established, it cannot be altered. He sees the naturalistic inquirer as one who views the program as being in constant flux, and is able to gain new insights as the program unfolds. This appears to be a false characterization. If a program changes in outlook and direction, the quantitative evaluator, too, can accommodate these changes by modifying the evaluation design, although with admitted difficulty.

Guba also sees the quantitative evaluator as one who intervenes and sets up the conditions for the independent variables to operate and the dependent variables to be measured. The qualitative evaluator observes the important interactions as they develop, as would a theater-goer attending a play. Guba fails to acknowledge the fact that a program is a form of intervention. It is not decreed by an evaluator, but is imposed on clients by the director and associated staff. A good evaluator would not impose conditions that are in conflict with program goals and objectives.

One other major dimension on which qualitative and quantitative evaluations differ, according to Guba, is that conventional inquiry involves a monolithic search for a single scientific truth, whereas naturalistic inquiry recognizes the esistence of multiple truths or perspectives. Guba is of the opinion that conventional evaluators naively believe that their investigations are free of bias and that, with the proper techniques, they can discover the single objective truth about the program or other entity being studies. Naturalistic inquirers, on the other hand, tend toward a pluralistic view; they seek multiple perspectives from program staff and clients regarding the program, while recognizing that these perspectives may contain bias. It is clear that any evaluator worth his salt would not fail to consider the viewpoints of program staff and clients.

A related issue mentioned by Cuba is that quantitative, as well as qualitative, evaluations may be contaminated by the prejudices and value systems of the evaluators themselves. Although this is certainly true, it can be argued that the techniques used by quantitative evaluators serve as a partial safeguard against such bias. Because the methods utilized are more structured and formalized, the quantitative evaluator is more likely to arrive at reproducible conclusions than is the qualitative evaluator.

Patton (1980), another advocate of qualitative evaluations, makes a distinction between qualitative and quantitative *data*. He states that "qualitative data consists of *detailed descriptions* of situations, events, people, interactions, and observed behaviors; direct quotations from people about their experiences, behaviors, attitudes, beliefs, and thoughts; and excerpts of entire passages from documents, correspondence, records, and case histories...The data are collected as open-ended narrative *without* attempting to fit program activities or people's experience into predetermined standardized categories such as the response choices that comprise typical questionnaires or tests... Quantitative measurement relies upon the use of instruments that provide a standardized framework in order to limit data collection to certain predetermined response or analysis categories...Qualitative measures describe the experiences of people in depth. The data are open-ended in order to find out what people's lives, experiences, and interactions mean to them in their own terms and in their natural settings. Qualitative measures permit the evaluation researcher to record and understand people in their own terms" (p. 21).

4. How Can Decision-makers Be Provided with Maximally Useful Information?

That descriptive data contain a richness of information about a program cannot be denied. In fact, a comprehensive description of the setting, the population, and the implementation of a program is an essential element of any good evaluation report. In order to be of maximum use to decision-makers, however, an evaluation must go beyond pure description. Some published evaluation reports focus intensively on the experiences of a small number of individuals, without indicating whether they are representative of the larger group. Although this sort of narrative often makes interesting reading, it does not provide the kind of information needed by program directors. When there are multiple interviews, procedures such as content analysis allow the evaluator to report the percentage of the people interviewed who held a particular opinion about the program. Some advocates of ethnographic approaches believe that by reporting percentages, one has somehow defiled the data. Certainly, the reporting of percentages or proportions does not rule out the inclusion of illustrative quotations from the interviews. Such quotations are informative and are useful in assessing the meanings of percentage figures. To report the responses verbatim without indicating what proportion of students make similar statements does not convey adequate information to decision makers. To maximize the reporting of useful information, a number of students should be interviewed and the data summarized in a table. For example, if a program involved 200 students, randomly select 25 to interview. Interview the students about their experiences in the program and tape-record their answers. Then examine each set of responses for content.

It should be mentioned at this point that content analysis is a quantitative technique. A set of mutually exclusive and exhaustive categories is developed, and the number of responses falling into each category is counted. The use of this post hoc

qualification in no way limits the response to the question. Some contributors to the evaluation literature fail to recognize that conventional quantitative techniques can be applied to open-ended responses.

With respect to the evaluation of achievement, the authors reiterate their contention that quantitative methods are most suitable. For example, if a program is designed to improve the arithmetic skills of sixth grade students, the most straightforward way to assess this goal is to examine school achievement as measured by test scores and answers to arithmetic problems based upon the program curriculum. This is the kind of information that the school board desires and it is the kind of information that the tax-paying community expects. Comparisons across social classes, ethnic groups, sex, and aptitude are certain to provide useful information that will aid decision makers. Comparisons are most easily performed with standard statistical models such as the analysis of variance, analysis of covariance, and multiple regression. Of course, the evaluation need not be limited to the assessment of arithmetic skills only. Attitudes, classroom environment, and any other variables of interest also could be examined.

To further clarify how evaluation studies should be conducted, a qualitative evaluation which has received considerable attention in the evaluation literature will be examined.

The TCITY Evaluation.

The evaluation of the Twin City Institute for Talented Youth (TCITY), was conducted by Robert Stake and Craig Gjerde in 1971, (Kraft, Smith, Pohland, Brauner, and Gjerde, 1974). This study is and example of a qualitative evaluation which could be greatly strengthened through the utilization of conventional quantitative techniques.

The Program.

The Twin City Institute for Talented Youth is a summer program for gifted youngsters from Minneapolis and St. Paul. The fifth session of the Institute met for seven weeks during the summer of 1971 on the campus of Macalester College in St. Paul. Talented youngsters in grades eight through eleven were eligible to apply. The report states that approximately 820 students attended. A total of 23 courses were offered, covering a wide range of subject areas; e.g., chemical bonding, Russian, graphic art, wilderness leadership, etc. Students were assigned to one of these courses based on their stated preferences. The school was staffed by 28 master teachers, 28 associate teachers, an unspecified number of teaching assistants, three administrators, and a secretary, as well as two evaluators.

Program Goals.

The specific objectives of the program are not clearly stated. The program director's statement of the goals, included in the report, ends with the following summary: "Having come interested in math, science, poetry, or archaeology, students will leave more committed to understanding and appreciating the total concerns of men. Success for the Institute will mean that students and teachers have infected each other with a personal honesty and will have demonstrated that learning can truly be humanizing" (p. 103). Another section of the report (p. 125) alludes to several other goals: (1) to provide an educational program that is challenging, stimulating, and relevant, (2) to provide master teachers with the highest ability to teach, (3) to provide younger teachers a good opportunity to learn more about teaching, and (4) to develop curriculum ideas that can be used in the public schools.

The Evaluation.

The evaluation report contains almost no information about how the evaluation was conducted. Stake, Gjerde, and their staff conducted classroom observations and collected questionnaire and interview data from students and teachers, a portion of which is incorporated into the report. In addition, their write-up includes "an advocate's report and an adversary's report, one summarizing the most favorable arguments in support of the Institute and the other summarizing the most damaging criticism. The reader is left with responsibility of resolving these conflicting arguments" (p. 100). It is claimed that the inclusion of this advocate-adversary report enhances the objectivity of the evaluation.

The main body of the report consists primarily of impressionistic descriptions of each course, based on classroom observations. The reader does not know (1) how many times each classroom was observed, (2) how the observations were scheduled, (3) whether teachers knew ahead of time when classes were to be observed, (4) whether more than one staff member observed each classroom, or (5) whether all observation reports were written by one person or submitted separately by individual observers. Each course description includes the course name, teacher name, number of students attending, and the stated aims of the course. Beyond this, the descriptions do not share a common structure. They vary in length from two sentences to two pages, and differ greatly in content. Some feature evaluative remarks about the teachers: "teachers personally sensitive but pedagogically insensitive" (p. 112); "Associate teacher, left in charge, gave us evaluators a hard time" (p. 114). Some make judgments about the students: "Students productive, somewhat too dependent, with a not unexpected fixation on sex-theme poetry" (p. 115). Still other reports are relatively neutral accounts of

the activities which took place.

With regard to the interview and questionnaire data, the report does not clearly state the number of measures administered, how they were designed, and how the sampling was carried out. The reader is left wondering whether those questioned were representative of students and teachers as a whole. Additionally, because only a portion of the data is included in the report, there is no way to determine whether these excerpts were representative of the data that were collected.

While acknowledging that the evaluation of this program must have been an awesome task, given the nebulous goals and the apparently small budget, the evaluation could have been improved and the information yield increased through the utilization of more systematic methods of collecting and summarizing the data.

An Alternative Approach to the TCITY Evaluation.

The first step in tackling this evaluation would be to establish a series of meetings with program administrators and teachers. The purpose of these meetings would be to (1) identify measurable goals and develop questions for investigation corresponding to each of these goals, (2) determine what kinds of data should be collected, when, and from whom, and (3) develop the data collection instruments to be used.

In practice, the authors would not limit themselves to examining only the stated goals and objectives. After studying the program, the investigation of additional constructs, might be recommended, based on knowledge of the literature, as well as previous research and evaluation studies. If, as the program progressed, some interesting but unanticipated effects became evident, these would be assessed as well.

It is of utmost importance in any program evaluation to solicit the assistance of the staff in the accomplishment of these tasks. Staff members are the best source of information on how the program is to be implemented, how to design appropriate instruments, and how to go about collecting data. Holding such meetings also gives staff an opportunity to meet the evaluators, ask questions about the evaluation, and to find out exactly what the evaluation will entail. In addition to working with the teaching staff, interviewing a number of students and seeking their assistance in identifying the questions to be investigated is helpful.

The presentation that follows focuses on the methodologies that should be employed.

Questions for Investigation.

As described above, a list of questions in cooperation with students, teachers, and administrators would be drawn up as follows:

(1) Did the students gain in their understanding of human values?

(2) Did students achieve the stated aims of the particular courses they were attending?

(3) Did students feel that TCITY was a positive experience?

(4) Did teachers feel that TCITY was a positive experience for their students and themselves?

(5) Did the manner in which classes were conducted reflect the values and goals of the Institute?

One aspect of the program which would not be assessed immediately is the briefly mentioned goal

of developing curriculum ideas for use in the public schools. This would be postponed until, perhaps, the middle of the following school year.

Although these questions seem quite straightforward, the original evaluation report does little toward answering them. Some information is provided regarding student satisfaction and, to a lesser degree, teacher satisfaction. The remaining questions, concerning the degree to which the Institute's goals were met, were not addressed.

Data Collection Procedures.

Next, a decision would be made about the types of data which would be needed in order to answer the above questions. Brief descriptions are given below of the measures proposed.

(1) *Humanistic Values Questionnaire (HVQ):* This measure would assess the degree to which the Institute met its general goal of fostering humanistic values. Admittedly, such a measure would be difficult to develop. However, because the fostering of humanistic values is apparently a major program goal, it should be assessed in some formal manner.

(2) *Class Aims Questionnaires (CAQ):* A separate questionnaire would be designed for each of the 23 courses, in order to determine the degree to which each met its specific aims.

(3) *Student Evaluation of Program (SEP):* This questionnaire would determine the students' level of satisfaction with the program.

(4) *Teacher Evaluation of Program (TEP):* This measure would assess the teacher's satisfaction with the program in terms of their own enrichment as well as the gains achieved by their

students.

(5) *Student interview:* A structured interview would be used to elicit more detailed information about students' opinions of the Institute.

(6) *Teacher interview:* A structured interview would also be administered to teachers to gather further information about their opinions of the program.

(7) *Classroom observations:* Using a structured observation checklist, evaluation staff would collect information about the classroom environment and the manner in which the program was being implemented.

Table 1 indicates when these measures would be administered. In addition, it would be important, and presumably simple, to collect demographic information about the students, such as age, sex, and ethnicity, as well as whether the student had been admitted under the Special Admissions program, whether he or she had attended the Institute before. Collecting such data provides a comprehensive description of the pupulation being studied. By collecting this demographic information, evaluators can assess the impact of the program on various subgroups of the total population and such information is useful in determining the exportability of a program. Assuming the current program was found to be successful, a school administrator planning to develop a similar program elsewhere would probably want to know how comparable the two student populations were.

Sampling.

Summarized below are the sampling procedures to be used for each instrument.

HVQ: It would be ideal to have responses from

all students on this measure. To avoid pretest sensitization, however, a table of random numbers would be used to select half the members of each course to take the pretest; the other half of the student body would take the posttest.

CAQ: The same sampling strategy would be used here. In this case, each course would receive a different questionnaire.

SEP: This questionnaire would be administered to all students.

TEP: This would be administered to all master teachers, associate teachers, and teaching assistants.

Student interview: These would be administered to a stratified random sample of the students. Each course would be represented in proportion to its size. This same approach also could be applied to the SEP if economic considerations prevented its administration to all students.

Classroom observations: Like interviewing, conducting classroom observations is costly. It would be imperative, however, to observe each class at least three times during the summer, utilizing two observers each time. Each observer would rate the classroom environment with respect to the variables deemed relevant. It could be stipulated that each class be rated at least once during each third of the summer session. It would also be important to observe each class on at least two different days of the week and during different portions of the class meeting.

Development of Measures.

Staff and students would be asked to identify potential items for use in questionnaires and in interviews and observation checklists. By using

structured interview and observation procedures, assessing the same set of variables would be assured across all units of data collection--teachers, students, or classrooms. Such "quantitative" procedures serve as a safeguard against the haphazard collection of odds and ends of information which characterizes the classroom observations conducted in the TCITY study.

After an item pool had been assembled, items would be selected and refined and tentative versions developed of all measures for review by staff. If possible, the measures would be piloted on small groups to assess their practical utility and psychometric adequacy.

Data Analysis and Presentation of Results.

For the HVQ and CAQ, data would be analyzed using standard statistical procedures such as tests and the analysis of variance in order to determine whether learning and attitude change had taken place during the course of the program. Also, in the case of the HVA, the demographic information that had been collected would be used in order to determine whether the degree of change was dependent on what grade the students were in, which classes they were attending, whether they had past experience in the program, etc. The major findings on the HVA and CAQ would be summarized in terms of averages and percentages and would provide detailed analyses on request.

In reporting the results of the SEP, the TEP, and the student and teacher interviews, descriptive statistics would be used. For each of the multiple choice questions, the proportions of individuals who gave a particular response would be indicated. All responses to open-ended questions would be subjected to a content analysis. The proportions would be reported, along with illustrative verbatim quotations. On the SEP, again demographic

information would be used to determine whether the degree of student satisfaction differed across arious sub-groups. The results of the classroom observations would be used to create individual profiles of each classroom, which could be included, in graph form, in the report.

In addition to the results obtained on these measures, the report would include a comprehensive description of the program, including its history and goals, the setting, staff, and students, and also the manner in which the program was conducted.

Summary.

In describing how to conduct the TCITY evaluation, the authors have had two major advantages over the original evaluators: (1) the benefits of hindsight, and (2) the luzury of discussing the evaluation without having to do it. Nevertheless, the methodological changes recommended are important ones. These changes can be summarized as follows:

(1) The employment of statistically sound sampling procedures, e.g., the use of true random sampling when it is not possible to assess all members of a group. This insures that results will be truly representative of the entire group of students or teachers.

(2) The collection and presentation of complete demographic information. This makes it possible to provide a comprehensive description of the student population and to assess the impact of the program on various sub-groups.

(3) The use of structured interviews and observations. This insures that all entities within a particular category (teachers, students, classrooms) are assessed on essentially the same dimensions and the use of a predetermined structure

reduces the amount of variance due to the interviewer or observer; that is, it increases the reliability of the measurement.

(4) The use of statistical techniques to examine change over time and to compare various sub-groups. Although statistical techniques are, of course, fallable, they permit more confidence to be placed on the findings than would purely subjective impressions.

(5) The inclusion in the report of all results obtained on questionnaire and interview measures, not merely excerpts based on unspecified criteria.

(6) The inclusion in the report of information about how the evaluation was conducted which gives the reader a framework for the interpretation of results.

An evaluator who utilized these quantitative techniques would produce information which would be of maximum utility to program administrators, staff, and clients.

REFERENCES CITED

Fienberg, S. E. The collection and analysis of ethnographic data in educational research. *Anthropology and education quarterly*, 1977, 8, 50-57.

Filstead, W. J. Qualitative methods: A needed perspective in evaluation research. In T. D. Cook and C. S. Reichardt (Eds.), Beverly Hills: Sage Publications, 1979.

Guba, E. G. Toward a methodology of naturalistic inquiry in educational evaluation. (Center for Study of Evaluation Monograph Series in Evaluation, U. C. L. A., 1978.

Kraft, R. H. P., Smith, L. M., Pohland, R. A., Brauner, C. J., & Gjerde, C. Four evaluation examples: Anthropological, economic, narrative, and portrayal. (American Educational Research Association Monograph Series on Curriculum Evaluation, Vol. 7.) Chicago: Rand McNally College Publishing Company, 1974.

Parlett, M. & Hamilton, D. Evaluation as illumination: A new approach to the study of innovatory programmes. In D. Hamilton, B. MacDonald, C. King, D. Jenkins, & M. Parlett (Eds.) *Beyond the Numbers Game.* Berkeley: McCutchan Publishing Corporation, 1977.

Patton, M. Q. *Qualitative Evaluation Methods.* Berkeley: Sage Publications, 1980.

Table 1. Instruments for the Evaluation and Times of Administration

Instrument	Beginning of Program	Middle of Program	End of Program
Humanistic Values Questionnaire	X		X
Class Aims Questionnaire	X		X
Student Evaluation of Program Questionnaire		X	X
Teacher Evaluation of Program Questionnaire		X	X
Student Interviews			X
Teacher Interviews			X
Classroom Observation Checklist	X	X	X

THE "METHODOLOGY" OF QUALITATIVE EVALUATION: THE CASE OF EDUCATIONAL CONNOISSEURSHIP AND EDUCATIONAL CRITICISM

Elliot W. Eisner

Stanford University

Precis: Elliot Eisner's unique point of view of evaluation is well documented in the research literature and in his book *The Educational Imagination: On the Design and Evaluation of School Programs*. In this article, he sets forth the methodology he uses in a clear, straightforward manner and illustrates its application in studies done by his doctoral students at Stanford University.

During the past decade there has been a rapidly growing interest in what is called "qualitative evaluation."[1] This interest is in part due to the realization that standardized achievement tests do not provide all of the information that educators want in order to know what teachers have taught and what students have learned. Partly as a result of the growing literature on the "hidden curriculum"[2] those in the field of education have recognized that students learn a great deal more than they are explicitly taught and that teachers and the school as an institution teach more than

either teachers or school administrators realize. Many of the most important lessons that students learn in schools are not those embodied in the explicit curriculum.[3]

But even more, educators working in schools have become increasingly conscious of something that football coaches, teachers of violin, and voice coaches have known for a long time: If you want to improve the way people perform, paying attention to the final score or to the performance on opening night is insufficient; it is important to see how the game is played during practice sessions and how rehearsals go before the curtain raises. Qualitative methods of evaluation are in large measure designed to focus upon the processes of educational practice in order to provide practitioners and others with information that cannot be secured from the scores that standardized achievement tests and other forms of summative evaluation provide.

The work that my students and I have been doing at Stanford University over the past decade is a species of "qualitative evaluation". It goes by the rubric of *educational connoisseurship and educational criticism* and it takes its leading ideas from the work that critics of the arts do when they appraise the efforts of writers, painters, film-makers, producers, actors, dancers, and others who make, direct, or perform works of art.[4] In this brief article I will describe the major features of this approach to educational evaluation, I will identify some of the concerns and issues that are commonly raised about it, and I will describe some of the methods that can be used by those who wish to employ educational criticism in their own work in schools and in universities.

Educational connoisseurship and educational criticism rest upon several assumptions. First, it is assumed that if individuals are to be able to describe the characteristics of complex social

situations such as classrooms they must be able to see the interactions among the characteristics that occur. Second, seeing such characteristics requires more than merely looking. It requires being able to discern what is significant given some frame of reference. In this sense seeing is regarded as an *achievement* that is cognitive in nature, it is not simply a task to be performed.[5] Third, it is assumed that such situations can be described in a variety of ways, for example, social situations can be described through the use of propositions, through number, through literary prose, by means of poetry, and through visual narratives on video, through cinematography, and by visual still photography. Fourth, it is assumed that each form of representation has unique utilities and limitations for revealing the significant qualities that one has seen. Fifth, it is assumed that the function of description through any of the forms in which it might occur is to help some person or group better understand the situation being described than he would without the benefit of the description. Thus, the function of critical narrative, regardless of the medium used, is to help people see, understand and appraise the character and quality of educational practice and its consequences.

What my students do is to observe classrooms for extended periods of time (from two to ten weeks), and write educational criticisms about what they have seen. Their ability to see what is subtle yet significant is a function of the level of educational connoisseurship that they possess and the presence of particular qualities of life and learning within the classrooms that they visit. For qualities to be seen two conditions must be present: The qualities must be in the situation in the first place and the individual observing must have the ability to note their presence.

The development of whatever levels of connoisseur-

ship students achieve is fostered through several means. Students enroll in a course in educational connoisseurship and educational criticism that I teach. In this course they read and discuss the criticism of professional arts critics and of students who have written educational criticism. They read and discuss doctoral dissertations that have used educational criticism as a dominant methodology, they observe, write and discuss the teaching practices of elementary and secondary school teachers they view on videotapes, and they visit classrooms and write educational criticism directly. These educational criticisms are then duplicated and discussed in class. The major aim of these exercises is to heighten their awareness to the subleties of teaching and learning. Discussion with peers about what each has seen is one way in which observations can be tested, refined, and modified. For those who wish to do doctoral dissertations using educational criticism this process might be extended over the course of the year, even before they visit the schools and classrooms they wish to study and write about.

Before discussing the structure of the writing used in educational criticism, it will be helpful to know something about the kinds of topics or problems that my students study. Thus far twelve doctoral dissertations have been completed at Stanford that use educational criticism as a cominant method. I shall briefly describe the focus of three such dissertations.

One study was designed to identify the differences in curriculum content that was taught by three elementary school teachers working with the same curriculum unit--the planetary system--within the same science curriculum.[6] The three teachers had about the same amount of classroom experience, they taught children of the same age and socio-economic status attending three schools within one school district. The major concern of the study

was to determine the similarities and differences in the content provided to students by these teachers. Put another way, the study asked: Given the same curriculum unit, what differences in content do children have access to as a function of the concerns, methods of teaching, and characteristics of the teacher? Given the fact that elementary school students spend a great deal of time each day with a single teacher, knowing something about how different teachers interpret content and manage student learning is important. Examining what students have learned about the planetary system through a common achievement test will not reveal what they have learned that was due to the unique contributions of their particular teacher.

A second study focused on the covert consequences of the use of a token economy to motivate elementary school students to learn mathematics.[7] This study combined both quantitative methods and educational criticism to determine the effects of different reward structures on mathematical achievement and other forms of learning.

A third study now in process is using educational criticism to identify the differences and similarities among three primary school teachers, each of whom has been identified as being very successful, moderately successful and unsuccessful in the teaching of reading as measured by their students reading achievement scores over a three year period and by the principal of the school in which they teach. The study is aimed at identifying and describing what accounts for the substantial differences in student performance. For example, it *might* be that the most successful teacher of reading achieves high student performance by using methods that might not be in the long-term best interests of the students. It *might* be that methods that are used are positive, subtle, and that once identified and described can be

replicated by other teachers with appropriate in-service education.

These studies are designed to provide information about teaching practices that neither summative evaluation procedures nor student observation schedules are likely to reveal. In all cases the students doing the studies have been elementary school teachers whose average tenure in the classroom is approximately eight years.

What is the character of writing in educational criticism? What is its structure? What are its salient features?

Educational criticism typically takes the form of a written document whose aim it is to help others see, understand, and appraise the quality of educational practice and its consequences. This document has three components. I describe these components separately for purposes of clarity, but in the course of the writing they may be integrated and difficult to separate into neat compartments.

The first aspect of educational criticism is descriptive. What the writer attempts to do is to provide a detailed and vivid description of the situation he or she wants to help others understand. The writing of descriptive criticism is both factual and artistic. The factual aspect of the writing might employ numbers, direct quotations of teachers' and students' discourse, or propositional descriptions of the physical characteristics of the situation about which the criticism is written. The artistic aspect of description is literary and metaphorical: indeed, it can even be poetic in places. The task of the writer is to communicate what is salient about the situation and what the teacher and the critic have agreed to focus upon. In order to optimize communication, the potential of language is exploited so that the literary and the factual complement each other. Much that needs

to be said about classrooms, teachers, and the quality of life and learning can be conveyed best not by factual description, but by literary narrative, metaphorical description, analogy. The problem of the writer is to help the reader vicariously participate in events that the reader has not experienced directly. To achieve this end the critic knows that emotions play an important role in understanding and knows how to use language so that it can reveal through emotion as well as through the presentation of the facts. When an educational critic[8] describes Dizzy Gillespie as a "middle-aged Tom Sawyer" when Gillespie performed for students at a San Francisco junior high school, the reader gets a sense of the man in a way that a factual description of his age, height, weight, race or behavior can never provide.

A second aspect of educational criticism is interpretive. What educational critics do is to account for the interactions they perceive in school situations. Why do these activities occur? What explains this particular episode? Why is it that this teacher is so effective? What is the nature of the "game" being played in this classroom? The effort here is not simply to describe the situation, but to interpret its processes and for this theory in the social sciences is appealed to. A good journalist might have the skills to vividly describe the situation but would not have the background to explain in theoretical terms the events that occurred. Doctoral students who have taken courses in educational psychology, sociology, history of education, and philosophy of education, are expected to draw upon concepts and theories from these disciplines to account for the events they have perceived and vividly described. This is far from an easy task: we tend in schools of education not to make the connections between theory and practice. But the connections are crucial if theory is to be more than an academic rite of passage instrumental to the achievement of

a higher degree. By attempting to make applications of theoretical ideas, both theory and practice may be improved. One might be able to see when theory falls short and where practice can be illuminated by theory that is appropriate to it.

A third aspect of educational criticism deals with the task of making value judgments about the educational merits of what have been described and interpreted. Evaluation is a valuative activity just as education itself is a normative one. What we want education as a process to accomplish is not merely to change students, but to improve the quality of their lives. We want schools to be *educational* institutions--rather than miseducational or non educational ones. To determine the *educational* value of the events perceived the educational critic appraises or evaluates these events by using criteria that are appropriate to their character. By this I mean that there are multiple views of educational virtue: there is no single "good" in education. The educational critic attempts to appraise the situation in its own terms. If the teacher is giving a lecture, the appropriate criteria to apply deal with the possibilities of the lecture as a medium of instruction, not the criteria appropriate to small group discussion, even if the critic himself prefers such an approach. Taking a leaf from the arts, if the play is a tragedy, we apply tragedic criteria to it, not comedic. If the school is a conventional "back-to-basics" school those criteria are applied, not the criteria appropriate for open schools. In short, within the limits of what is regarded as educationally sound and the varieties in which such soundness may appear, the criteria to be applied to appraise a classroom or a school is to be suited to the "species" of educational practice described, interpreted and appraised.

In the end a written document will be prepared and educational criticism that is descriptive,

interpretative and evaluative will be provided to some audience. The particular way in which the criticism is written depends upon the audience for whom it is intended. Since different audiences have different needs and different aims, and since the goal of educational criticism is to contribute to the improvement of practice, the criticism's relation to the audience is a crucial consideration. If the audience is a board or funding agency that wants to know how the project that it funded four months is doing, the criticism is written a a way suitable for such a concern. If the criticism is for the teacher who wants to know about how he or she is relating to students, how he or she is presenting ideas or leading discussions in class, then the criticism can be written with these needs in mind. Not everything that might have been observed needs to be shared at one time. Some feedback might be in a verbal form, in what tone, depends upon the critic's assessment of the situation and what is likely to be constructive.

I mentioned that the nature of the message is related to the needs of the audience. This relates as well to the critic's focus. When an agency, foundation, or at times a teacher requests or agrees to having an educational criticism prepared, the focus of the criticism is often *prefigured* that is, the focus is determined by the specific needs of those groups or individuals. But the focus of an educational criticism might also be *emergent* A critic might be invited to a school or classroom without a prefigured focus and after several days or weeks perceive an aspect of the school or classroom that is of considerable significance but which could not have been anticipated. For example, one of my students received permission from a secondary school English teacher to observe and to write an educational criticism about her class. What emerged during my student's observation was the extraordinary way in which the

teacher used satire in her teaching. It was not the case that the teacher was herself teaching satire; it was that she was satirical in her teaching.[9] Such a process or an approach could hardly have been prefigured. There is no research on teaching that deals with the use of satire, nothing on this topic in the *Handbook of Research on Teaching*, or in the *Education Index*. The point here is that the focus of criticism can be either or both prefigured as a part of the research bargain between the critic and the teacher or it can be emergent. Indeed, in general, the latter is more prevelant in my students' work than the former.

What are some of the issues and concerns that people raise about such an iconoclastic approach to educational evaluation? Those professionally socialized in the tradition of the social sciences are concerned with matters such as (1) the reliability of observations, (2) validity and (3) the generalizability of the findings or conclusions. The most frequently raised concern deals with reliability. How does one know whether there is bias in perception and reporting? The question itself reflects naievete concerning the inevitable "bias" in any form of observation. Any observation schedule or test leaves as much or more out than it includes, regardless of the form. In this sense, any form of data collection is biased by the assumptions on which it was built and the methods that it employs. Pristine objectivity is an epistemological impossibility. The empty mind sees nothing.

Reliability, or perhaps a better term, believeability, is achieved through a process of structural corroboration, a process something like "triangulation" in sociological methodology, or circumstantial evidence in law. The writer of criticism provides compelling conclusions by supporting those conclusions with evidence

embodied within the criticism that relates to events which have transpired. This process is much like what all of us do in our daily lives. We put pieces together as they emerge and we form a whole that makes sense to us. We do not typically engage in a controlled experimentation or use blind judging procedures in order to make inferences about the future or to draw conclusions about the present.

A second manner in which reliability can be determined is by assessing what is called the *referential adequacy* of the criticism. Good criticism should illuminate what it addresses. An educational criticism that is reliable *and* valid will enable someone with less connoisseurship than the critic to see what otherwise would have gone unseen. Just as a good music critic can help us hear counterpoint unheard, melodic lines unrecognized, the test of the educational critic is to be able through the critic's work to perceive what is subtle and complex within classrooms and schools. The educational critic is something of a midwife to perception.

As for generalizability, this is not a major concern for those doing evaluation but it is a major concern for those who are engaged in educational research. According to conventional research canons, generalizations can occur only if the samples researched have been randomly selected from a universe whose parameters have been defined. One can generalize if one knows these parameters and if the process has been random. Putting aside the fact that such assumptions are honored more in the breach than in the fact, the belief that random selection is the only valid way to generalize is itself questionable. In fact, it hardly squares with our ordinary experience; we form the generalizations we hold seldom through a process of random selection. Generalizations occur by a build-up of common general features

recognized in particulars. What we learn from particular situations we use to form expectations for the future. We modify those expectations as a result of subsequent experience. Educational criticism, by helping us perceive, understand, and appraise particular situations also therefore helps us form generalizations about educational practice. If we were restricted to practice in education only what we had learned from studies in which random selection was employed, we would be compelled to close the school house door tomorrow. Thankfully, we learn a great deal from a wide variety of methods of inquiry and in forms that are far wider than what can be stated in formal propositions or through number.

There are, of course, a variety of issues that could be addressed in the analysis and appraisal of educational criticism. Space does not permit a full treatment here and hence I refer the reader to the works on educational criticism listed in the references. There is, though, one point that I would like to make about the efficiency of educational connoisseurship and educational criticism as a method of educational evaluation. Educational criticism, compared with the use of standardized observation schedules is not an efficient method. It takes time. It takes subtlety of perception. It takes considerable skill in writing. It requires the ability to apply theory to practice. It requires one to make educational judgments. Further, there is no recipe--like procedure or "method" comparable to computing a test for doing this work. It is, in short, a method that requires no small degree of artistry. To know what educational criticism provides one must read it. It is more complex than looking a a set of raw scores, T-scores, percentiles, or means. In this sense, it not only takes time to write, it takes time to read as well. Yet, I know of no better way to learn how the game has been played than to watch it with sensitivity

and intelligence while it is in process. Without knowledge of this process I know of no credible way to improve what we do in the schools. Evaluation as a process can perform many different functions in education: it can reward, it can screen, it can select, it can at times help us determine if our objectives have been achieved. But perhaps the most important function of evaluation from a strictly educational point of view is to help educators improve the quality of educational life for students. There are no short-cuts that I know of to this end. Educational connoisseurship and edcuational criticism have a modest contribution to make to its realization.

REFERENCES CITED

1. The word qualitative evaluation, as distinct from quantitative evaluation is the result of a conceptual mistake. All empirical research is rooted in the description of qualities. Number is one mode of description. See my forthcoming article: "On the Differences Between Artistic and Scientific Approaches to Educational Research." The Educational Researcher, forthcoming, April, 1981.

2. For an interesting article on the hidden curriculum see Elizabeth Vallance, "Hiding the Hidden Curriculum," Curriculum Theory Network, Vol. 4, No. 1, 1973-74, pp.5-22.

3. In my book The Educational Imagination I distinguish between three types of curricula, the explicit, the implicit and the null curriculum. Although the explicit curriculum is the one that gets the most attention from educators, the other two may be equally --perhaps even more important.

4. For a full discussion of these concepts see Elliot W. Eisner, The Educational Imagination:

On the Design and Evaluation of Educational Programs, Macmillan and Company, 1979.

5. This distinction between an achievement verb and a task verb is made by Gilbert Ryle, The Concept of Mind, New York: Barnes and Noble, 1949.

6. Gail McCutcheon, "The Disclosure of Classroom Life," Doctoral Dissertation, Stanford University, 1976.

7. Betty Sternberg, "What Do Tokens and Trophies Teach". Doctoral Dissertation, Stanford University, 1977.

8. Robbie Schlosser, "A Musical Program," mimeo, Stanford University School of Education.

9. Ann Sherman "Humor and Satire," mimeo, Stanford University School of Education.

Some doctoral dissertations that have used educational criticism as a dominant methodology are as follows:

Robin R. Alexander, "Educational Criticism of Three Art History Classes," Stanford University, 1977.

Thomas Barone, "Inquiry Into Classroom experience: A Qualitative, Holistic Approach," Stanford University, 1978.

John F. Feilders, "Action and Reaction: The Job of an Urban School Superintendent," Stanford University, 1979.

Elizabeth Vallance, "Aesthetic Criticism and Curriculum Development," Stanford University, 1975.

Also see Elliot W. Eisner, "On the Use of Qualitative Evaluation for the Improvement of Teaching," Journal of Educational Evaluation and Policy Analysis, Vol. 1, No. 6, December 1980.

THE CALIFORNIA STAFF DEVELOPMENT STUDY

Bruce Joyce and Robert N. Bush

Precis: Bruce Joyce of Booksend Laboratories of Palo Alto, and a former Director of Teacher Education at Columbia University and the University of Chicago, and Bob Bush, the well-known professor emeritus of Stanford and Director of the Laboratory for Research and Development in Teacher Education, now are the co-directors of the California Staff Development Study. This important statewide evaluation is an excellent example of the use of qualitative and quantitative methods with two of the best scholars in education in charge.

The California Staff Development Study is a longitudinal four-and-a-half year study which commenced in the fall of 1980 under a contract between the State Department of Education (SDE) and Booksend Laboratories of Palo Alto, California. The two major purposes of the study are: (1) to provide periodic status reports on the implementation and mpact of the various federal, state and local staff development initiatives which provide opportunities for educational personnel to learn and apply new skills; and (2) to provide on an outgoing basis feedback of study results to a variety of individuals with an interest or stake in staff development,

from the Legislature to school faculty members and administrators.

The first study purpose reflects s summative evaluation emphasis. Information gathered through an annual questionnaire survey, an analysis of the implementation of legislative initiatives, interviews, and case studies will, by the end of the study, allow an assessment of the overall merits of the initiatives and the impact they have had in the classroom.

The second study purpose reflects a formative evaluation emphasis. That is, information will be fed back to Legislators, officials in the State Department of Education, county officials, school district administrators, directors of teacher centers, teacher and administrators in schools, and others who provide facilitative service to schools. The design concentrates on finding out what information each of these groups needs and encouraging collaboration within and among these groups to use this information to effect improvement in the delivery of staff development programs. Thus, at the local level, teachers and administrators will be asked about their information needs, to be trained by the contractor to participate in collecting data relevant to those needs, and periodically receive results of the investigation to aid in the local improvement of programs. The formative feedback and training function is designed to enhance the capacity of local educators to plan for and continuously improve their own competence.

The impetus for the California Staff Development Study comes from the fact that staff development represents a significant opportunity for improvement of both the basic school program and the various categorical programs. Given the fact that substantial amounts of funds flow into programs with staff development components in them, state

legislative and executive policy makers want information to help assess the overall merits of these programs as well as to refine and improve program policies. Until now, there has been no comprehensive study of the various initiatives, the staff development opportunities offered through these initiatives to school personnel, and the education practices resulting from taking advantage of these opportunities. Without this knowledge base, significant improvement in the coordination and delivery of staff development is unlikely.

Preliminary work for the study has been conducted over the past two years. In 1978, the SDE's Office of Program Evaluation and Research (OPER) commissioned a group of individuals to design a strategy or studying staff development in California. In 1979, OPER let another contract for the design of instrumentation and data collection procedures for this study. Using this preparatory work as a conceptual base for a study, OPER awarded a contract to Booksend Laboratories to conduct the longitudinal California Staff Development Study commencing in September, 1980.

The study will be conducted in two phases. During the first two and a half years the primary purpose will be to provide information about the state of the art of staff development in California and to feed that information back to personnel in schools, districts, counties, the State Department and the Legislature in order that the initiatives at all levels can be improved and effectively coordinated. The second phase will be a continuation of that effort but with the additional objective of attempting to determine the effects of various kinds of staff development on teacher behavior and on pupil behavior and achievement. As a picture of staff development emerges during Phase I, the impact of staff development on children's learning will be explored in a set of investigations designed to increase knowledge about the kinds of activities

which most effectively help teachers increase their competence and the effect of increased competence on student learning.

The study has six major components or tasks:

- an analysis of legislative initiatives from conception through implementation
- an annual questionnaire survey
- regular interviews about staff development with personnel at all levels
- in-depth case studies in several clusters of schools, including a study of both the formal and informal systems of local staff development and the training of local data collectors
- formative feedback to personnel at all levels levels
- special policy studies

The Analysis of Initiatives

In California, over 40 state and federal initiatives with staff development as either their primary focus or as a significant component within them have been identified as operating at the district or school level. These initiatives can be classified into four categories:

(1) Agency-creating or strengthening efforts (e.g., the federal Teacher Centers serve regions or school clusters)

(2) Direct resource provisions (e.g., AB 65 provides resources directly to schools)

(3) Curriculum-oriented projects

(e.g., the California Writing Project serves teachers of writing)

(4) Population-oriented programs (e.g., bilingual education, special education)

The initiatives vary greatly in how they operate or are administered at the local level. Some, such as AB 65 and AB 551, provide resources on a regional basis and thus have more variable linkages to schools. The initiatives are described more fully in *Recommendations for the Evaluation of Staff Development in California: Report of a Preparatory Study*. (California State Department of Education, 1979).

The intent of the analysis of initiatives is to make available a comprehensive, compact, and usable summary of the major staff development initiatives. These initiatives will be analyzed in terms of their goals, their administrative structures, their implementation patterns and costs, and their target audiences.

The Annual Questionnaire Survey

Each year, teachers, administrators, and community members in 200 schools in eight counties will complete questionnaires for the study. The questionnaires will provide quantitative data regarding demographic characteristics of the schools, the nature, frequency, and utility of the more formal staff development activities at the school site, perceptions of the more pressing problems and suggested solutions. These data will be extremely useful in identifying those initiatives which appear to operate at the school level, and in providing meaningful benchmarks on the status of staff development in the state.

Interviews

In each of the counties a network of teachers and administrators will be developed and regularly interviewed to provide information about needs and issues. These interviews will supplement the questionnaire survey and will illuminate with practical detail the major issues, problems, and successes of staff development in the state.

Case Studies

During the first year, in-depth case studies will be conducted in several clusters of schools to explore the dynamics of the schools and the extent to which the formal and informal systems impact upon the manner in which school personnel improve their competence. The interaction of the initiatives with the social ecology of the schools and the ability of teachers to influence their own development and the quality of their professional lives will be of central importance. Teams of local data collectors will be trained to conduct the case studies and provide feedback to their peers. If successful, this training component will serve as the prototype for an expanded component in subsequent years. Training in the conduct of case studies will be made available to a much larger population.

Formative Feedback

Formative feedback activities in this study will be planned with the help of program participants at all levels of the delivery system: state, regional, district, school, and community. Feedback will be given on results of the questionnaires, analysis of initiatives, interviews, and case studies, and will be responsive to the following types of questions: (1) What is the nature, extent, and perceived effectiveness of staff development activities offered at the relevant level of feedback?;

(2) How well do these activities meet the needs of participants?; and (3) What could be done to improve the situation?

Special Policy Studies

Provision has been made in the study for the State to commission Booksend Laboratories to conduct short-term policy studies on particular topics that may be of current concern to the State. Each special policy study will draw upon the basic data generated by the California Staff Development Study, and will also collect additional data bearing upon the particular policy problem of interest. The first such policy study commissioned is in the area of special education staff development for regular classroom teachers, and will be completed in early 1981.

MELE-MAAT

James C. Stone

At the invitation of the South Pacific commission, my wife and I spent two weeks in June, 1980 at Mele-Maat, a British primary school (1-6 grades, called classes) serving the rural villages of Mele and Maat in the New Hebrides. The school had been selected as a typical school by the South Pacific Commission at which to demonstrate the effects of a special program of in-service education under their direction. In the New Hebbides, English is the medium of instruction and children are taught English, beginning in class (grade) 1. This study had as its main function the collection of baseline data for a three year evaluation.

The investigation was conducted during a period of critical national crisis as the British, French, and New Hebrides governments sought to prevent a takeover by secessionist French-speaking rebels prior to New Hebrides independence on July 30, 1980. The New Hebrides government, which took over after UN-sponsored elections in November, 1979, is English-speaking. This independent archipelago now is named Vanuatu.

This report is a "Readers Digest" version of the complete study which is available from the author. In the summarized version, we begin with some necessary background information and the evaluation design as an example of the use of qualitative and quantitative data; we move on to findings regarding reading, writing, and speaking competence; and we conclude with our observations about teacher and pupil attitudes toward school and English as a second language.

1. Background of the Problem

The 74-year history of British-French rule of the New Hebrides is a history of two separate groups vying for authority and partisans. Most clearly this has been reflected in the educational system. Or, we should say systems, for there are primary-secondary schools and a training college for teachers where *all* instruction is given in French and there are primary-secondary schools and a training college for teachers where *all* instruction is in English. In both instances, the children also speak their own mother dialect. When it is time to enroll in first grade, parents must choose to send their children either to a school where all instruction is in English or French.

Schooling. Mele-Maat, the focal point of this baseline evaluation, is a rural British primary school (1-6 grades) serving the villages of Mele and Maat. The school year is divided into three terms--February-May, June-August, September-December. Children go to school five days a week, Monday- Friday from 7:00 a.m. until 11:30 and from 1:30 to 3:00 p.m. Children and teachers go home for lunch. Instruction follows a similar schedule for all grades and includes a variety of lesson content (oral English, reading, writing, math, science, physical education) for 10-15 minutes each (classes 1 and 2) to 20-30 minutes each

(classes 5 and 6). The afternoons offer some variation from the established routine with periods of arts and crafts, agriculture, singing, and religious instruction.

The district government selects and pays teachers, allocates funds for construction of classrooms, and provides textbooks. The local community provides the school site, constructs the classrooms, builds the furniture, and purchases equipment, supplies, and materials. The villages of Mele and Maat raise the money for these items and the principal deposits the funds with the district and draws on them as purchases are made. The schools are not free--parents pay a fee (estimated at about $15.00 U.S. per year at Mele-Maat), and provide uniforms that are worn by all the children. An advisory group of parents (called the School Committee) advises the principal. In the New Hebrides, approximately three-fourths of the primary-age children attend school.

Mele-Maat, like other primary schools, was originally a mission school and currently has an enrollment of approximately 30 pupils per class in six classes. It was extremely rundown until three years ago when the present headmaster was appointed. He was successful in obtaining broad community support and today Mele-Maat compares favorably with other rural village schools in the New Hebrides.

The Faculty. Information regarding the teachers is as follows:

Class	Name	Age	Sex	Approx. No. of Years of Teaching Experience
1	Touas Sanford	38	F	15
2	Asel Poilapa	34	F	7
3	John Shing	43	M	10
4	Edwin Ngaga	55	M	25

5	Allen Bule	33	M	8
6	Selwyn Arutangair	21	M	3

All are natives of the New Hebrides and obtained a secondary school education there and a year of teacher training. The fifth class teacher also is the headmaster.

The Teaching of English as A Second Language.

In order to appreciate the competitive and controversial milieu of the New Hebrides, it is important to understand that the language of the school--English vs. French-- is the center of the conflict between competing English and French systems of education. The children coming to school from the village of Mele speak their own mother tongue. Similarly, those from the Maat village speak their own. (The two groups communicate with each other through Bislamar, a pigeon English, that is the *lingua franca*) Once they begin class 1, however, all instruction is in English.

Learning a second language in the formal, artificial, structured environment of a classroom is different from learning a first language informally, naturally, functionally. With English as a child's only language, the school emphasizes the irregular, the unusual, the exceptions, the more subtle nuances, based on the premise that language learned naturally has provided the basic structure. In ESL, teachers emphasize regular, uncomplicated, common language syntax, structure, and usuage. Those familiar with teaching *naturally*-learned languages need to understand the limitations that are imposed upon a school-learned *foreign* language. The fact that practice occurs only a few hours a day at the maximum in a British primary school is a serious limitation to ESL achievement in listening, speaking, reading, and writing. Thus, it is inappropriate to use as the basis of

comparison of Mele-Maat student achievement, the level of English gained by students whose schooling is all first language learning. For the purpose of our evaluation, therefore, we chose to compare Mele-Maat student achievement with those in comparable schools serving similar populations where English is taught as the foreign tongue.

The SPC/Tate Language Program.

In the Mele-Maat School, English as a second language is taught through the system developed by G. M. Tate and published, disseminated, and implemented throughout the Pacific by the SPC. The Tate syllabus is based on a structured framework and is designed to provide an integrated approach to the skills of listening, speaking, reading, and writing for pupils learning English as a foreign language. The total package comprises 15 Oral English syllabi. Written English comes from the reading course and later from oral language development. The program is used by most South Pacific Commission-served countries. For the past 15 years, in addition to consultant services, the SPC has provided in-service education to teachers utilizing this program.

Of the current situation, the Commission's English language specialist wrote:

> "We are now in the position that the Tate materials, in part, have been used in the Pacific for some time. *They have never been evaluated.* We are at a point today where the validity of the English programme is questioned, partly because it does not seem to provide the vehicle for the weighty content deemed to be necessary for primary school children, and partly because of the low achievement of secondary school children."

Faced with these vexing conditions and the fact that

the SPC/Tate Language program never had been evaluated except by subjective selective remembrance of headmasters, the SPC decided to sponsor this three-year evaluation.

2. The Evaluation Design

The purpose of the evaluation is to determine the effectiveness of the teaching of English and instruction in English as a second language in the Mele-Maat School over a three-year period.

Questions to be Investigated and Means of Assessment

1. What is the school like as a school and what is the school climate, *as judged by on-site observations and photographic analysis of the facilities?*

2. What is the quality of the teaching of ESL and the quality of English as the medium of instruction, *as assessed by school site observations and classroom recordings that are evaluated by a panel of ESL specialists?*

3. What is the level of pupil reading and listening achievement, *as measured by the Micronesian Achievement Test Series?*

4. What is the level of pupil and teacher writing ability, *as measured by writing samples that are evaluated by a panel of Bay Area Writing Project specialists?*[1]

5. What is the level of pupil and teacher oral competence *as measured by recorded samples of classes as well as selected individuals and that are evaluated by a panel of Performance Evaluation in Foreign Language specialists?*[2]

6. What is the attitude of the children toward

learning English and instruction in English, *as appraised by interviews, case studies, and observations?*

7. What is the attitude of the teachers toward pupils learning ESL, the use of English as the medium of instruction, and themselves as teachers, *as appraised by interviews, questionnaires, and observations?*

8. What is the attitude of parents toward their children learning ESL, the teaching of ESL, and the use of English as the medium of instruction, *as appraised by interviews and observations?*

The Research Strategy.

The two basic research strategies to be used in this evaluation are triangulation and replication. Triangulation, i.e., securing the same data from multiple sources or from the same source by three different methods, ensures cross-validation of findings. For example, interviews, questionnaires, and case studies were used to obtain similar information from the same population; at least three independent judges rated written and oral samples of pupils' and teachers' English, and three independent observers judged school site adequacy.

Replication will be used by repeating the study in 1981 and 1982. Replication also ensures cross-validation of findings.

The Obtrusive Observer.

No matter how hard the observer tries to be unobtrusive, the moment he walks onto the school grounds or enters the classroom, he contaminates the situation and what is observed is not typical, having been changed, however subtly, by the outsider's presence. This obviously was the case at Mele-Maat. White Americans, serving as external evaluators, were intruders into the normal routine

of the school for a two-week period. Just as the student teacher enlivens his teaching (and the class itself perks up) when the University supervisor is present in the classroom, so the observers were aware that the Mele-Maat teachers (and their pupils) sometimes contrived to look their best for the evaluators, despite all assurances that "looking good" was unimportant now, and only important a year or two hence.

Validity.

We were acutely aware of the perplexities of establishing rapport as well as other threats to validity. Repeatedly we reassured the teachers of our purpose--to survey the present as a basis for what is to follow. Often we went out of our way to emphasize the positive, complimenting the teachers on lessons that were well-taught whenever we observed them. We mingled informally with teachers and pupils at breaktime, noontime, activity periods, on the playground, and after school. Gradually, the teachers became less apprehensive, more communicative and trusting. Gradually the pupils became less reserved, more responsive and informal. In the end, we were convinced we had achieved a reasonable amount of truth in what we saw, heard, and recorded.

In any evaluation, those doing it are viewed as a threat by those upon whom it is being done. Especially is this true when the evaluators are external and foreigners at that! Consequently, the rapport we finally were able to establish with the children was particularly gratifying. They especially loved our tape recorder and whenever we recorded their classwork, we played it back so they could hear themselves. This pleased them immensely. They giggled, laughed, and pointed to each other as they listened. We became good friends, and even later going into classrooms to observe without the tape recorder, we received a

joyous and cordial welcome accompanied by clapping of hands as we sat down, and waves of goodby when we left.

3. School Facilities and School Climate

Physical Facilities. The Mele-Maat School comprises four buildings forming an L around a grassy field. *Grades 1 and 2* are housed in separate buildings, each a patchword construction of tin, concrete block, boards and woven bark-like material. Light is poor and ventilation inadequate. In both classes, the children sit two to a wooden desk. There is a teacher's table and a cupboard in each room as well as a blackboard.

Grades 3 and 4 are together in another small building. A wall separates the two classrooms, but not the noise which often is disturbing and distracting. The same is true of *grades 5 and 6* the only difference being that the latter two classrooms, instead of being separated by a wall, have the school "library" between them. The "library" is 8 x 30 feet with shelves along one wall containing the textbooks and other reading materials. In reality, it is a storeroom.

Both the 3-4 and 5-6 classroom buildings are substantially constructed and they are tidy and neat compared to the smaller buildings of classes 1 and 2. The two buildings housing classes 3 and 4, 5 and 6 are well ventilated and the light provided by windows along two sides of each room is quite adequate on sunny days. There is no electricity in any of the buildings and on cloudy days, the lighting is poor.

Equipment, Books, Supplies, and Materials.

Equipment consisted of two sewing machines, saws, soccer balls, volleyballs, and rubber balls. Books, at the rate of 1 per 2 1/2 children, were

in use. Copies not being used were stored in the library. The quality of the textbooks and workbooks was good, but there were no Teachers' Manuals.

Supplies were meager. The available writing paper was scrap; sewing materials were remnants. Each pupil had a pencil and workbook and rulers were available.

School Climate.

On the playground the children played well together and got along well. In classes, the children were well-disciplined, even quite shy and reticent by U. S. standards. Teacher and pupil morale was high. There was great pride in the school and what it has become in three years and the achievement of its graduates in the past two years.

Dramatic evidence of the positive school climate and the supportive attitude of parents was provided at the farewell parent-teacher luncheon given us on our last day. The headmaster had told us the parents wanted to have tea with us, following the School Assembly. The tea turned out to be a complete luncheon, beautifully served on plates brought from the village and featuring native foods, along with cookies and cakes the women had baked. In attendance were a dozen parents as well as the faculty and their spouses. During the luncheon, Edwin, the 4th class teacher who was the spokesman for the group, proudly showed us pictures of himself with the crew of a World War II U. S. battleship when he was 14 years old! The group's enthusiasm and friendliness were touching.

4. ESL Teaching and English as the Medium of Instruction

Teaching was traditional. Most of the instructional day was spent in teacher talk, with children

responding on command, usually in unison and by rote. Occasionally, individual children were asked to respond. All blackboard work was done by the teacher. The teacher typically was "up front and on high" with a pointer, rarely moving around the room. With the exception of pre-reading and reading instruction in classes 1 and 2, the teaching observed was whole-class instruction with all pupils expected to proceed at the same pace. In pupil responses to oral English questions/commands, or in other subject areas using oral answering via rote memory, some pupils often did not participate. Those who did made up for the others by responding loudly! The quality of textbooks and workbooks was good, but they were rigidly followed in an "all-together and at-the-same-time" fashion. Quicker students had to wait for others to catch up and the slowest ones who did not understand English simply were ignored.

ESL.

As stated earlier, ESL was done by rote, the teacher giving directions, followed by pupil responses in unison, much repetition, a few individual responses, and teacher corrections of pronunciation. Typically, complete sentences were used. Teachers seldom varied from the textbook or used the lesson to create analogies or illicit informal responses. Volunteers usually responded and only occasionally were non-volunteers called on.

Written work usually was limited to copying sentences or paragraphs; seldom was writing of the children's own. A good example was our request for a creative sample of writing a paragraph, an incident, or a short story of the children's own making. We explained this carefully, emphasizing that we wanted to see the present status of their writing ability as a prelude to seeing some improvement when we returned next year. We thought our intent had been made clear, yet the next day

the samples turned in by two of the teachers all were copies of the same written paragraph.

In reading, we observed pupils reading together in sing-song. Typically, the teachers read first and then the pupils read. This was done sentence by sentence, seldom by paragraphs. No silent reading was observed, nor children reading aloud without the teacher first reading the passage for them to imitate. No oral reading of stories to the class by the teacher was observed.

We found no evidence that English was spoken at home. Since it was not used typically by children on the playground, we tend to doubt it.

Tyranny of the SPC/Tate System in Practice.

Teachers typically used one book per term, following it to the letter day-by-day. Writing was attached to the reading program after the first three oral books and continued up to and including paragraphs via workbooks. Copying sentences and paragraphs was much in vogue. Oral competence was seen as being the single, all-important goal.

The SPC/Tate syllabus, however, was not intended to be followed slavishly and repetitiously. The idea was to illustrate situations so that teachers (and pupils) could create analogies based on their own experiences. But, since there were no teachers' manuals at Mele-Maat, the students' text-books also were the teacher's guide.

A maxim in acquiring language competence is that free-time reading habits need to be developed. Generally people who write well also read a lot. The implication is clear: if the teachers and pupils are to improve their use of English, opportunities and encouragements must be offered for free-time reading. We saw only a few books

that could be taken out by pupils for this purpose. As we said earlier, the only books available for teachers were the students' textbooks. We do not know if they read other books at home, but we tend to doubt it. The New Hebrides do not have a reading cultural tradition and home-reading in English usually is limited to newspapers and magazines and home-listening to the radio.

In summary, English is a captive of the classroom. This was true of the children. Of greater import, it also was true of the teachers. And therein lies the problem.

5. Reading and Listening Achievement

The Measuring ESL Reading and Listening Achievement Tests were administered to 135 children in classes 2-6. These are the silent reading comprehension and listening to oral reading sections of the Micronesian Achievement Test Series (MATS). A summary of the reading comprehension scores are shown on Table 1.

As shown in Table 1, in Class 2, most of the pupils have not yet learned to read. The highest raw score was 14 of 50. Marked improvement was achieved by Class 3 with half of the pupils achieving a raw score of 33-42. Class 4 had the highest achievement with 10 pupils having a raw score of 44-49. In Class 5, 10 students achieved a raw score of 40-46 and in Class 6, 11 students achieved raw scores of 37-44. In every class, except Class 2, there were from 2-10 non-readers, i.e., scored ≤ 17.

The reading tests were developed within the confines of the SPC language course and test achievement only within that framework. They are not convertible to standard reading ages and are of limited application. They will, however, provide a base for comparison in subsequent years. The

Table 1

Summary of Reading Comprehension Scores: Mean, Standard Deviation, Range, and Ages of Mele-Maat Pupils, by Class.

Class	N	Reading Comprehension			Age	
		Mean	Standard Deviation	Range	Mean	Range
2	21	3.2	3.6	0-14	8.0	6.1-9.1
3	20	30.3	9.3	3-42	9.3	7.8-12.4
4	33	33.6	11.6	9-49	10.3	8.5-12.3
5	32	26.5	14.2	0-46	11.8	9.9-14.0
6	29	33.5	7.6	12-44	12.3	10.4-14.4

same is true of the Listening Tests. In the meantime, other appropriate tests will be sought for use in subsequent years so that a comparative base can be established between reading and listening achievement of Mele-Maat pupils and other ESL children in developing nations.

6. Pupil and Teacher Writing Competence

The Mele-Maat teachers were requested to provide a sample of each pupils' free-writing. The samples of teachers' writing were their responses to an open-ended questionnaire. All samples were holistically evaluated by staff of the Bay Area Writing Project.

Pupils.

On the whole the more spontaneous writing--as in classes 5 and 6--showed a great deal of difficulty with forms of English verbs, subject-verb agreement, and the use of singular-plural distinctions. Some pupils also had trouble with articles. Vocabulary, perhaps because of their limited experience and background, was quite restricted; fishing, gardening, a few school matters, the family, and help around the house were the subjects. In summary, all of the pupils would benefit from:

-hearing a wider range of English spoken by native speakers and speaking with them.

-Being read to from good books in English that could expand their limited vocabularies, then trying to use those new words in stories of their own.

Teachers. Generally, the teachers were fluent; some tense and usage errors were apparent but these are relatively minor. Class 5 and 6 teachers performed better than the others in their

writing. Classes 1-4 teachers had more subject-verb and number agreement problems, more spelling errors, and more usage and English idiom problems. They were less fluent than those of class 5 and 6 teachers, but their writing was understandable. In summary, the teachers would benefit from:

-reading a wide range of books in English in their free time that could both expand their vocabularies and sensitize them to good syntax and good literature.

-writing stories, essays, anecdotes in English and joining together with an English teacher or editor to critique and revise their writings.

7. Pupil and Teacher Oral Competence

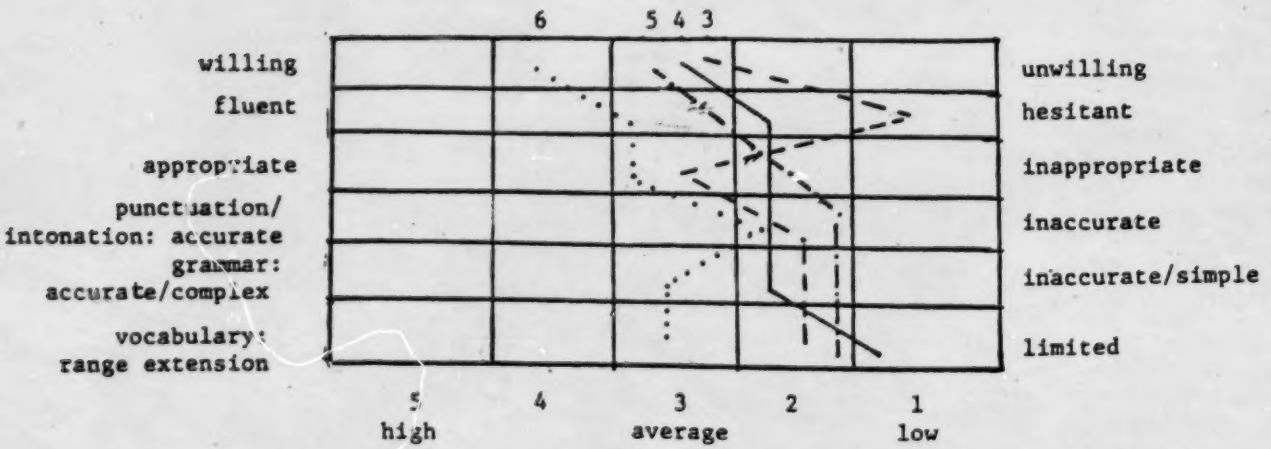
In addition to observing and listening to the English spoken by pupils and teachers, we tape-recorded lessons in oral English as well as lessons in the various subject fields taught in the Mele-Maat school. We also recorded informal conversations between ourselves and the pupils, between pupils and other pupils, between teachers and ourselves, and teachers and other teachers.

We analyzed the tapes and independently arrived at ratings. We then asked staff of the PEFL project to evaluate them, using the Harding Oral Test Profile.

The standard used by all raters was English as a Second Language. Three kinds of ratings were secured: one for pupils whose interviews had been taped (two each from classes 3-6), one for teachers using English as the medium of instruction in the teaching of maths, science, geography, and other subjects in the Mele-Maat curriculum. The pupil rating results are shown in Table 2. The two sets

TABLE 2

The Harding Oral Test Profile: Pupil Interviews

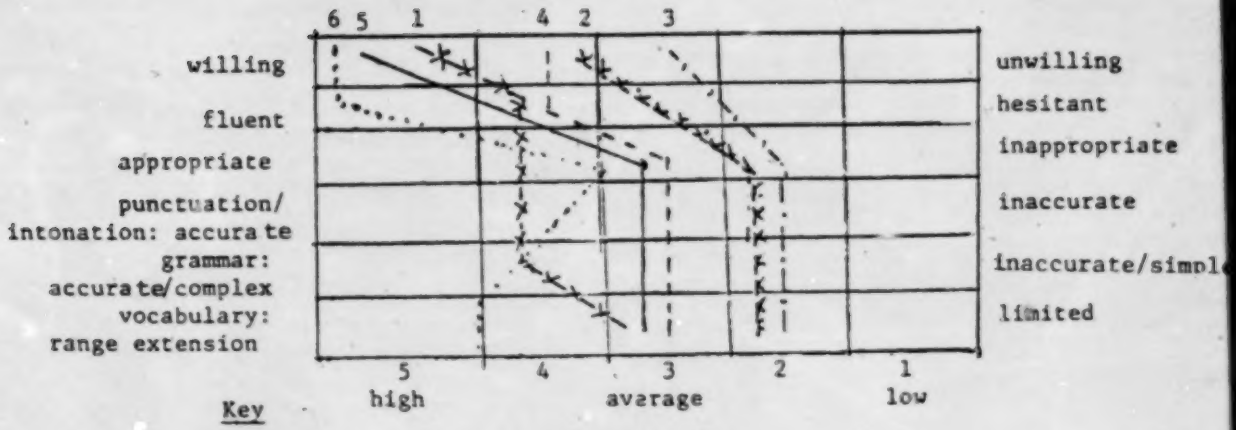


120

TABLE 3

The Harding Oral Test Profile: Teacher Interviews

Global Response



Key

..... represents the 6th Class Teacher

———— represents the 5th Class Teacher

-------- represents the 4th Class Teacher

—·—·—·— represents the 3rd Class Teacher

xxxxxxx represents the 2nd Class Teacher

x-x-x-x represents the 1st Class Teacher

of teacher ratings were combined and are shown as consensus ratings in Table 3.

In general, pupils' oral competence increased with school class, as expected, although we found some younger children who clearly were superior to many older ones. In general, the teachers divided into two distinct ability groups, with teachers of classes 6, 5, and 1 more competent than those of 2, 3, and 4. This range in quality between the first, fifth, and sixth class teachers and those teachers of classes 2, 3, and 4 was particularly significant when related to the pupils writing samples. There is a direct relationship between high quality oral competence of teachers and their pupils' oral and writen competence.

The results of combining consensus teacher ratings were:

willingness	---average to high
fluency	---low to average
appropriateness	---high
pronunciation	---low to average
vocabulary	---low to average

Tables 2 and 3 also show a correspondence of higher marks on the first three categories (willingness, fluency, and appropriateness) and lower marks in the last three categories (punctuation, grammar, vocabulary). This was true of both pupils and teachers. Generally there was little use of dipthongs in pronunciation. Typically, teachers and pupils cut short their pronunciation, as indicated in stress and intonation, thus tending to render the output monosyllabic. We determined that the general understanding level of teachers to be average to high. On the whole, most of the teacher's oral expression was understandable. There were, however, common repeated mistakes. A few of the most commonly heard errors were:

Tense mixing:

Ex. Using present when past is proper; using present and past in the same sentence; using incorrect verb forms-- "It is impossible to have everybody goes to...", "You are awake from sleep", "What are they do/ What did they doing", "She is give them."

Confusing singular/plural forms of nouns/verbs:

Ex. Two pencil. One pencils. Children differs/ childrens differ. "She saw seven chair." "Impossible to have everybody goes."

Word endings:

Ex. ng not pronounced in "ing" endings, "sing" for singing, "s" for "d" as in clothes for closed.".

Ignoring consonant blends:

Ex. "dun" for "don't"
"th" for "t" as in "tick" for "thick"
"ch" for "sh" as in "chells" for "shells"

Vowel miscueing:

Ex. "al" for "il" as in "pensal" for "pencil"
"al" for "aw" as in "aways" for always"
"oo" for "o" as in "flaw" for "floor"

Lacking auxiliary:

Ex. "How many right angles has a window? for "How many right angles does a window have?"

No use of Bislamar was observed in the classroom. Occasionally it was used on the playground. English was seldom used by pupils or teachers in informal

conversation among themselves or on the playground.

Some Suggestions: The teaching strategy of the Mele-Maat teachers should be modified if the results of their anticipated increased competence is to impact directly on their pupils. The standard rote pattern repetition form of teaching should give way to teaching on the basis of spontaneous conversation in small groups. This is the method that is at the heart of the new approach to foreign language being implemented by the PEFL Project.

An example of the PEFL system applied to the Mele-Maat School would be first for small groups of pupils to listen for comprehension to tapes of native speakers and these in turn would become the basis for informal, spontaneous conversations within the groups. The conversations would be taped for listening and self-evaluation while the teacher moved among the groups to critique and correct. Other groups would be listening to taped recordings of stories being read by native speakers and/or listening to an English radio broadcast.

Clearly the key to both better pupil and teacher oral and written competence is the teacher's oral ability. They need more models in the form of English-speaking visitors, greater informal intercourse among English language speaking people, tapes of oral expression to record and listen to themselves, listening to English via the radio at school and at home, plus direct oral language in-service education.

8. Teacher Attitudes

A six-item open-ended questionnaire was completed by the teachers to obtain information on their attitude and feelings about the school, the pupils, themselves as teachers, and their profession. These same questions were used in interviews with

them. A summary of the questionnaire/interview data follows.

1. Teaching English As a Second Language

All six teachers wrote about the difficulty of teaching English as a second language, the heavy responsibility they carried as models for the children, and the time and care required in preparation for oral work. Three mentioned the importance of and their dependence on the Tate materials, and two spoke of the necessity for in-service education to help improve their teaching as well as their own speech patterns. One quote from the class 6 teacher spoke for the group:

> "Teaching English as a second language is a very hard thing to do because one has to be sure of himself as a good English speaker and writer....I try my best to speak good English, especially during lesson times and outside the classroom in conversation with the children."

Two teachers cited the difficulty of teaching English in terms of their own native tongue and Bislamar. The fifth class teacher's comment was:

> "The difficulty comes from the fact that three languages are spoken by the children:
> first, their mother tongue;
> second, Bislamar;
> third, English.
> Bislamar is the one that makes English difficult to teach because many words in Bislamar sound like English. What happens is that English has tenses to show the verbs whereas Bislamar has no tenses--everything is in the present tense."

The class 2 teacher remarked how her teacher training had aided her in teaching English:

> "We've been taught how to teach it, the way lessons should be planned, the aids to be used, the new words or topics to be introduced, and then the drill and repeating in sentences."

Despite the problems, the difficulties, and their own admitted inadequacies, the teachers were unanimous in their opinion that they enjoyed teaching English and were eager to improve it as well as their own speaking and writing skills. The class 4 teacher succinctly said:

> "The children are very interested in their English learning. They like it. Not only in 'Oral English Periods,' but in Maths and other subjects as well--and the more they are eager to learn, the more the teacher has to help. It helps to make the teacher work hard."

The class 3 teacher cited the pleasure he received from teaching English by seeing his pupils able to:

> "...use English with visiting tourists almost every week, with parents at homes, (and) the fearlessness of the children with outsiders."

2. Things They Like About the Mele-Maat School

Surprising to the evaluators, who are familiar with the tradition of school uniforms from their experiences in a number of other developing countries that formerly were British, when asked "What do you like best about the school?", the teachers' number one response was "the school uniforms." Apparently, it has taken some "doing" to accomplish this and they were very proud of it.

Next, five of the teachers commented on the reforms at the school since 1978, citing the leadership of the headmaster with the community, the children, and the staff, and pointing to the progress made in having four students in 1978 and nine in 1979 pass

the Senior Primary Examination. One said:

> "When the parents saw the school progressing, they were and are still happy about it and are waking up to pay their children's fees whereas before we had to remind them again and again but this time the fees are going on very well. The children have a very good environment now than many of the schools in the new Hebrides."

Another major plus was the contribution of the SPC. The second grade teacher wrote the following about it:

> "The furniture, for example classes 1 and 2 now have the good desks and seats to sit on, like these from last year's work for the children always sits on the cold floor. Also the new cupboards to have whereas before the books usually be all over the places, then the building itself and the locks, nothing get lost now."

Improved deportment by the children and cooperation among the staff was discussed by two teachers in the following manner:

> "Children's manners and politeness, their cooperation with teachers, e.g., what he thinks of his teachers, etc."

The other teacher said:

> "I believe the teachers are cooperating together very well."

Three of the six teachers talked extensively about the SPC project--the added materials, the forthcoming inservice education program, and the three-year evaluation effort--as major assets of teaching at Mele-Maat. Some direct quotes provide a taste

of the intensity of their feelings.

> "Best in this school most of all is this language project, which to be carried out. Already from the project we have a good cupboards, extra books will be given by SPC to give enough books to the children and also teachers will improve their method of teaching English."
>
> "Personally, if what Miss Purdy and Mr. and Mrs. Stone are after is running smoothly--then our future will be much more happier than it is now."

3. How Do You Feel About Being a Teacher?

Perhaps they divined we wanted to hear it, more likely because they believed it themselves, all six teachers cited the many plusses they found in being teachers. Some typical comments were:

> "I like being a teacher because I was trained for the work."

Another:

> "My interest has been always teaching. I've tried many times to do different jobs but never happy as in teaching. I like teaching because I should share what I got. I also enjoy working with children and parents...this is why I feel teaching is a very interesting job for me. Therefore as long as I can work I want to remain a teacher."

Another:

> "Being a teacher there is a lot to do and learn. Being a teacher I know more about each individual child's character in class and his parents and his home. Being a teacher seems

to enjoy my life in school with both children and other teachers. I always find new things to teach to the class and myself."

Still another:

"I like being a teacher because chance to prepare children for their futures:
(a) At home in family life
(b) On their land as farmers
(c) On the water as fisherman
(d) In administrative work as officers, principals, teachers, etc.
(e) In hospitals as doctors, nurses, dressers, etc.
(f) In the country as artists, governments, etc."

Finally, two teachers coupled the satisfaction with the amount of hard work that was required of teachers. The second class teacher wrote:

"When I first started teaching, I thought to myself, oh, being a teacher is really a great to do beforehand. A lot of thinking to think about the work or the lesson itself/what to do. How to go about it, when to teach it and so on. .. I like being as a teacher, but also a lot of work to do with the children. *I like people to critique my work so that in this way I might know where I'm to learn from my mistake."* (Italics ours.)

The sixth class teacher summed up the attitude of all when he said:

"I think teaching is one of the hardest jobs. One may get to be honest. I felt as resigning the first year teaching but...it's one of the important jobs one can give to a developing country and to like any job, one has to get

himself into it...The thing that makes me feeling really interested is when I saw the results my children produced in their exams. I think teaching is one of the hardest jobs but also one of the interesting jobs and I like it."

In summary, the characteristics of the Mele-Maat staff that stood out were a positive attitude towards teaching and their dedication and hard work. They expend a tremendous amount of energy in a long and strenuous teaching day and we concur in their feeling that it is hard work but more than worth the doing.

9. Pupil Attitudes

In order to obtain more detailed and personalized information about pupils' attitudes toward school and learning ESL, we jointly interviewed two pupils from each class. Naturally we sought the advice of the headmaster and he suggested that each teacher select two pupils to be interviewed according to whatever criteria he/she wished to use.

The students whom the teachers selected turned out to be their best students. In informal conversation, these children were so shy, the evaluators were disappointed with the results of their interviews. We then asked Edward, the 4th class teacher who was an important person in the village of Mele to recruit someone to be an interviewer/interpreter for us. The next day, he brought a man who had been a teacher and, using our interview guide, he talked with the same eight children. They were equally shy with him and gave him essentially the same limited information we had obtained. If anything, our initial interviews provided more data. We felt it was important that a native speaker confirm this information we initially had obtained and it was reassuring to realize our rapport with the children had been at least as good as his, probably better.

Our complete report contains eight case studies. Reproduced here, however, is one case study and a report of the more dramatic School Assembly.

Kalisto is a small, slight, handsome boy, age 10, who is in the sixth class. His family consists of 8: grandfather, grandmother, mother, father, three sisters and himself. He is the oldest. His best school subjects are Math and Reading, in which subjects he says he does very well. He enjoys going to school and likes and admires his teachers. After school, he works in the garden, tending yams, manioc and cabbages; he also hunts in the bush and fishes. His favorite sports are football and high-jumping. At the end of class 6, he hopes to go to secondary school. When he grows up, he would like to be a farmer.

As revealed by our eight case studies, the pupils can be described as interested in learning, pleased to be learning English and satisfied to be taught in English, proud of their school, compatible with each other, shy and somewhat reticent in conversation about their family, friends, and in and out of school activities; friendly, responsive, enthusiastic.

The School Assembly

The students' positive attitude towards the school and their teachers was clearly evident in the pride and enthusiasm they showed at the School Assembly. The assembly was called to bid us farewell on the last day of our stay. Fortunately we have a complete recording of it, for the Assembly was an emotional event for us and the children. Their singing was beautiful and their wholehearted enthusiasm, calling out, and clapping were very moving. We presented a World Atlas to the school and the headmaster responded with appropriate remarks. As we left to the outpourings of goodbyes and applause, one of the 6th class pupils

gave us a beautiful shell he had discovered on a coral reef. He could have sold it for several dollars in town. Instead he bought our hearts.

The complete report concludes with a summary of the major findings of this formative evaluation, ten recommendations for improving the school, and a precursor of what the school might become by 1985.